# Essays in Econometrics:

Author: Joseph Cooprider

Persistent link: http://hdl.handle.net/2345/bc-ir:108714

This work is posted on eScholarship@BC,
Boston College University Libraries.

Boston College Electronic Thesis or Dissertation, 2020

Copyright is held by the author, with all rights reserved, unless otherwise noted.

# Essays in Econometrics

Joseph Cooprider

A dissertation

submitted to the Faculty of

the department of Economics

in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Boston College

Morrissey College of Arts and Sciences

Graduate School

March 2020

© Copyright 2020 Joseph Cooprider

**Essays in Econometrics**

Joseph Cooprider

Advisor: Arthur Lewbel, Ph.D.

## Abstract

In my doctoral research, I developed econometric estimators with strong applications in analysis of heterogeneous consumer demand. The first chapter develops an estimator for grouped patterns of heterogeneity in an approximately sparse setting. This setting is used to estimate demand shocks, competition sets and own-price elasticities for different groups of consumers. The second chapter, which is joint work with Stefan Hoderlein and Alexander Meister, develops a nonparametric estimator of the marginal effects in a panel data even if there are only a small number of time periods. This is used to estimate the heterogeneous marginal effects of increasing income on consumption of junk food. The third chapter, which is joint work with Stefan Hoderlein and Solvejg Wewal, is the first difference-in-differences model for binary choice outcome variables when treatment effects are heterogeneous. We apply this estimator to examine the heterogeneous effects of a soda tax.

**Chapter 1:** "*Approximately Sparse Models and Methods with Grouped Patterns of Heterogeneity with an Application to Consumer Demand*" introduces post-Lasso methods to time-varying grouped patterns of heterogeneity in linear panel data models with heterogeneous coefficients. Group membership is left unrestricted and the model is approximately sparse, meaning the conditional expectation of the variables given the covariates can be well-approximated by a subset of the variables whose identities may be unknown. I estimate the parameters of the model using a grouped fixed-effects estimator that minimizes

a post-Lasso least-squares criterion with respect to all possible groupings of the cross-sectional units. I provide conditions under which the estimator is consistent as both dimensions of the panel tend to infinity and provide inference methods. Under reasonable assumptions, applying this estimator to a consumer demand application allows me to partition consumers into groups, deal with price endogeneity without instrumental variables, estimate demand shocks, and identify compliments and substitutes for each group. I then use this estimator to estimate demand for soda by identifying different groups' competition sets as well as demand shocks using Homescan data.

**Chapter 2:** In "*A Panel Data Estimator for the Distribution and Quantiles of Marginal Effects in Nonlinear Structural Models with an Application to the Demand for Junk Food*", we propose a framework to estimate the distribution of marginal effects in a general class of structural models that allow for arbitrary smooth nonlinearities, high dimensional heterogeneity, and unrestricted correlation between the persistent components of this heterogeneity and all covariates. The main idea is to form a derivative dependent variable using two periods of the panel, and use differences in outcome variables of nearby subpopulations to obtain the distribution of marginal effects. We establish constructive nonparametric identification for the population of "stayers" (Chamberlain, 1982), and show generic non-identification for the "movers". We propose natural semiparametric sample counterparts estimators, and establish that they achieve the optimal (minimax) rate. Moreover, we analyze their behavior through a Monte-Carlo study, and showcase the importance of allowing for nonlinearities and correlated heterogeneity through an application to demand for junk food. In this application, we establish profound differences in marginal income effects between poor and wealthy households, which may partially explain health issues faced by the less privileged population.

**Chapter 3:** In "*A Binary Choice Difference-in-Differences Model with Het-*

*erogeneous Treatment Effects and an Application on Soda Taxes*", we answer how should Differences-in-Differences be implemented when outcomes are binary and we expect heterogeneous effects. The scope for applications is clearly vast, including labor force participation, product purchase decisions, enrollment in health insurance and much more. However, assumptions necessary to measure heterogeneous effects in classic Difference-in-Difference models break down with a binary dependent variable. We propose a model with a nonparametric random coefficient formulation that allows for heterogeneous treatment effects with a binary dependent variable. We provide identification of the average treatment effect on the treated (ATT) along with identification of the joint distribution of the actual and counterfactual latent outcome variable in the treatment group which allows us to show the heterogenous treatment effects. We suggest an estimator for the treatment effects and evaluate its finite sample properties with the help of Monte Carlo simulations. We further provide extensions that allow for more flexible empirical applications, such as including covariates. We apply our estimator to analyze the effect of a soft drink tax on consumer's likelihood to consume soda and find heterogeneous effects. The tax reduced the likelihood of consumption for the most consumers but not for those who were most likely to be consuming previously.

# Contents

# Chapter 1

# Approximately Sparse Models and Methods with Grouped Patterns of Heterogeneity with an Application to Consumer Demand

## 1.1 Introduction

Unobserved heterogeneity is a consistent issue in many microeconometric models. Even in large panel data sets with many variables, there is still large unexplained variation. Applied researchers face a trade-off between using flexible approaches to model unobserved heterogeneity and building parsimonious specifications. This heterogeneity can enter a model through slope-heterogeneity as well as unobserved individual and time heterogeneity.

Often researchers use individual or time fixed effects to capture much of the unobserved heterogeneity. These can often be biased by an incidental parameter's problem and often don't have enough data to accurately estimate. Further, these estimates often lead to computational difficulties because of large numbers of parameters. One way around this issue is to model individual heterogeneity as a small number of unobserved types (Keane and Wolpin, 1997; Hahn and Moon, 2010). These models may lead to the "incidental parameter bias" when dealing with short panels (Nickel, 1981). I will improve this by using a framework that allows for clustered time patterns of unobserved heterogeneity that are common within groups of individuals (Bonhomme and Manresa, 2015).

A further issue depends on slope heterogeneity. Traditional panel data models simply assume slope homogeneity such that the regression parameters are the same across individuals. While convenient and simple, this approach has been rejected in multiple studies (Hsiao and Tahmiscioglu, 1997; Lee et al., 1997; Durlauf et al., 2001; Browning et al., 2007). Other researchers will use individual specific slope heterogeneity where regression parameters are estimated for each individuals (Baltagi et al., 2000; Hsiao and Pesaran, 2008). This requires sufficient within individual variation of the covariates as well as a sufficiently long time-series component of the data. Further, these estimates could have high variance making inference difficult. These problems are exacerbated in cases where the number of parameters are large.

The basic model I use to address these challenges will take the following form based on the work of Bonhomme and Manresa (2015):

$$y_{it} = x_{it}'\theta_{g_i} + \alpha_{g_i t} + v_{it} \tag{1.1.1}$$

such that $i = 1, ..., N$ and $t = 1, ..., T$ where $N$ is the number of individuals and $T$ is the number of time periods. The covariates $x_{it}$ are contemporaneoustly uncorrelated with $v_{it}$ but may be correlated with the group-specifice unobservables $\alpha_{g_i t}$. The number of groups is set by the researcher. My addition to the work of Bonhomme and Manresa is that I assume that $\theta_{g_i}$ varies between groups and is approximately sparse, which means that the conditional expectation of the $y_{it}$ given $x_{it}$ can be well-approximated by a subset of the variables whose identities may be unknown. This can be seen as adding constraints such that the individual slope parameters are constrained to be equal to certain group parameters, and some of the group parameters are constrained to be zero.

This model uses group-specific time effects to address unobserved heterogeneity. This allows for time effects to differ across individuals while allowing larger portions of the data to identify each parameter which avoids many of the problems with standard fixed-effects approaches. In order to address the slope heterogeneity concerns, I use group-specific slope parameters and assume approximate sparsity. I will use Post-Lasso methods based on Belloni et al. (2012), which is an extension on the classic Lasso (Tibshirani, 1996), in order to shrink the parameter space which will improve estimation.

I will use this approach in a consumer demand framework. Consumer demand can be challenging because researcher have many heterogeneous consumers and many different prices. In a product space estimation approach, as the number of products increases, the parameters needed to estimate the model increases exponentially. Thus, in order to get stable estimation, researchers often make significant assumptions about consumer behavior. One way to do that is to impose functional form assumptions on utility and grouping products together. This is called the Almost Ideal Demand System (Deaton, 1980). Another solution to these problems is to use a product characteristic approach instead, but the researcher will still have to determine what characteristics to include as well as making other assumptions on consumer preferences. My approach will allow the data to tell us information about the consumer preferences and minimize the assumptions made by the researcher.

Assume $y_{it}$ is a log quantity and $x_{it}$ is log price. This would imply $\theta_{g_i}$ is the group specific own- and cross-price elasticities. Using my estimator allows each group to have different own- and cross-price elasticities and will set many of the cross-price elasticities to zero. Assuming that the cross-price elasticity space is approximately sparse is reasonable since consumers will often only consider a subset of all possible items when making purchasing decisions because of search costs, and may only consider the prices of some of those products, which I will call their competition set. My model is imposing constraints on price elasticities such that individual price elasticitis are equal to each other and some price elasticities are set to zero. However, this model allows the data to set which price elasticities are applied to, rather than leaving that to assumptions determined by the researcher's intuition.

In addition, these elasticities can allow us to identify different products' interaction effects (substitutes or complements) even when the number of products is large. It will also identify how some products may be substitutes for some consumers, while not (or perhaps even complements) for others. Identifying complements and substitutes from a large number of products is a helpful addition to the literature.

My model further informs us more about consumer demand by using time-varying group fixed effects. These time-varying fixed effects can help us identify different trends in demand between different groups of individuals. For example, consumers may not consume unhealthy food at the start of the year as a new-years resolution, but will eat more unhealthy food throughout the year and their resolution fades (Cherchye et al., 2017). There also may be seasonal trends in consumption that may not hold for everyone. Grouped time-varying fixed effects can then measure demand patterns and shocks for groups of individuals.

Further, these time-varying grouped fixed-effects will capture many of the shocks that would lead to endogeneity issues in my estimates. However, if a demand shock impacts the whole population or only subset of my sample, these shocks will be captured by $\alpha$ and they will not bias my estimates of $\theta$. Thus, I argue that I do not need to use instrumental variables to measure cross-price effects.

For my application of consumer demand, I will look at the Nielsen Scanner Data which is available through the Kilts Center at the University of Chicago Booth School of Business. I will focus on aggregate monthly purchases in 2014. I will estimate the demand for the most popular soft drink product based on demographic characteristics of the individual as well as the prices of the soft drink product and other popular soft drink products based on my data.

### 1.1.1 Related Work

Beyond the basic AIDS approach (Deaton, 1980) to consumer demand, another product space demand estimation was done by Hausman et al. (1994) in a three stage approach. First, they estimated demand for product category. Second, they estimated demand for the various groups of products and third was the demand for individual products within the groups. There has been additional work to help use the data to help identify which groups to use and include (Blundell and Robin, 2000). Adding heterogeneity into an AIDS was done by Jorgenson (1990). For a review of methods to add heterogeneity when aggregating results, see Stoker (1993).

A different possible demand estimation solution in this setting involves including many product characteristics to a BLP, (Berry et al., 1995a), type model. This is done by using lasso to estimate high dimensional product characteristics space by Gillen et al. (2014) and using double-Lasso (Gillen et al., 2015) to control for the large number of demographic characteristics. They can include many product characteristics, but considering how difficult it is to get product characteristics data, it is impossible to get every product characteristics consumers might consider.

Additional research addressing consumer demand with many prices and consumer heterogeneity using machine learning techniques has been done by Chernozhukov et al. (2019). To estimate the aggregate price and income-elasticities, they similarly use a Lasso estimate but rely on double machine learning to relax the sparsity assumptions (Chernozhukov et al., 2018). Heterogeneity is introduced on an individual level with individual slope parameters which are averaged (as in Chamberlain (1982, 1992)) and regulated using a ridge penalty for how far away the individual parameter is from the average.

While this structure allows a relaxing of the sparsity assumption, to model heterogeneity, it requires a long time dimension for identification and assumes time homogeneity of preferences. This assumption is common and useful in econometric panel data models (See Chernozhukov et al. (2013), Graham and Powell (2012), Hoderlein and White (2012), Chernozhukov et al. (2015) and Chernozhukov et al. (2019) for examples). While this assumption is helpful in allowing enough within individual varition to identify the slope parameters, it may not hold in consumer demand settings and is not required for my model.

There is a significant literature on estimating the subsets of goods a consumer considers when making a purchasing decision. Estimating the probability of each subset of goods being in an individual's consumption set has been of interest in psychology and marketing (Hauser and Wernerfelt, 1990; Shocker et al., 1991). In economics, they are often used as a generalization of discrete choice models (Caplin and Dean, 2015; Abaluck and Adams, 2017). Chiong and Shum (2018) estimate a discrete choice model with high dimensional choice sets using machine learning to shrink the initial product space to a smaller subset while keeping the estimation consistent. Barseghyan et al. (2019) establishes a model that does not require strict assumptions to estimate demand preferences and perform welfare analysis but is unable to point identify the choice sets. While my estimation can choose what I call an individual's competition set, I cannot identify their consideration set. There may be additional research on how these two sets compare.

Some papers in the economic literature have used individual choice data to esitmate parameters to estimate interactive effects (substitutes or complements) for a very small number of items; see Chintagunta and Nair (2011) and Berry et al. (2014) for summaries of this literature. Some researchers have attempted to treat each potential bundle as a discrete alternative and impose a parametric structure to identify complementary between items (Train et al., 1987; Gentzkow, 2007). Song and Chintagunta (2007) build a model to estimate whether to purchase and how much to purchase each good. However, because of the computational difficulties of these models, the number of goods included is very small and chosen by the researcher. My approach allows the data to tell us which goods may be compliments or substitutes.

There is also previous research dealing with heterogeneity and lasso using a random coefficients model (Fan and Li, 2012). While this allows some coefficients to be set to zero, it does not consistently let some variables set to zero for some individuals but not others and is in a cross-sectional setting. It also doesn't identify the parameters for each individual but instead identifies the distribution of the parameters.

Grouping individuals together is an effective way to measure heterogeneity when estimating consumer demand. For example, consumer brand choice analysis using individuals or groups of individuals yield similar general trends (Oliveira-Castro et al., 2006). Further, it is common to group individuals together based on where they live (Huang and Lin, 2007), their income level (Aasnass and Rødseth, 1983), their search costs (Koulayev, 2009), or some combination of these (Asano and Fiuza, 2015; Bester and Hansen, 2016).

Su et al. (2016) develops the C-Lasso to shrinks individual coefficients to group-specific coefficients. While it allows for individual fixed-effects, this model solely relies on the coefficients for group classification and does not group on unobserved heterogeneity. This is important in my setting sense I am grouping consumers not only on their price elasticites but their demand shocks over time. Further, Su et al. does not use a Lasso within each group, which often leads to large variation in situations with many covariates. The extenstion (Su and Ju, 2017) allows for fixed effects with time interactions based on Bai (2009) but often requires large time series data points for proper classification.

Ando and Bai (2016) expands on these ideas to use interactive fixed effects and minimizes sum of least squared errors with a shrinkage penalty to allow for large number of covariates. This paper differs from mine in a few important ways. It requires $N$ and $T$ to go to infinity simultaneously since it uses the SCAD penalty of Fan and Li (2001) and interactive fixed-effects. It also does not have a way to estimate the regularization parameter since cross-validation is normally infeasible because computation is very intensive in these types of models.

The number of groups in this paper and the papers listed above is fixed. There are methods in a standard panel model to use a *kmeans* clustering algorithm to identify the number of groups and sort the individuals into the groups prior to estimating the model (Bonhomme et al., 2017) which allows the number of groups to grow as the sample size grows. This does not allow for covariates to have different effects on each group which is a significant feature of my model. A different approach to this answer is done by Su et al. (2019) where individuals can change groups over time. My approach does not currently allow for such generalizations so this remains left for future research.

## 1.2 Estimator

This section will begin with a discussion of the model and estimator. Then I will discuss computational methods.

### 1.2.1 Model

This model (1.1.1), based on the model in Bonhomme and Manresa (2015), contains three types of parameters: the parameter vector $\theta_g \in \Theta$ for all $g \in \{1, ..., G\}$; the group-specific time effects $\alpha_{g_i t} \in \mathcal{A}$, for all $g$ and all $t \in \{1, ..., T\}$; and the group membership variables $g_i$ for all $i \in \{1, ..., N\}$ which maps individual units into groups. The parameter spaces $\Theta$ and $\mathcal{A}$ are subsets of $\mathbb{R}^P$ and $\mathbb{R}$, respectively. Let $\gamma \in \Gamma_G$ denote a particular grouping (or partition) of the $N$ units, where $\Gamma_G$ is the set of all possible groupings.

The grouped fixed-effects estimator is defined as the solution of the following minimization problem:

$$\left(\widehat{\theta}^{GFE}, \widehat{\alpha}^{GFE}, \widehat{\gamma}^{GFE}\right) = \underset{(\theta,\alpha,\gamma)\in\Theta^G\times\mathcal{A}^{GT}\times\Gamma_G}{\arg\min} \sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - x_{it}'\theta_{g_i} - \alpha_{g_i t})^2 \quad (1.2.1)$$

which is the minimum over all possible groupings $\gamma$ of the $N$ units into $G$ groups, common group parameters $\theta$, and group-specific time effects $\alpha$.

In my setting, which differs from Bonhomme and Manresa's model here, I will assume that the $\theta$s will be approximately sparse, defined formally below.

**Condition 1.** - *Approximately Sparse: Each optimal function of $f_g(x_{it})$ is well-approximated by a function of unknown $s > 1$ variables:*

$$\begin{aligned} y_{it} = f_g(x_{it}) + v_{it}, \quad f_g(x_{it}) = x_{it}'\theta_g^{AS} + \alpha_{g_i t} + a(x_{it}) \\ ||\theta_g^{AS}||_0 \leq s = o(N), \qquad \mathbb{E}_N[a(x_{it})^2]^{1/2} \leq c_s \leq \sqrt{s/N} \end{aligned} \quad (1.2.2)$$

This requires that there are at most $s$ terms that are able to approximate the conditional expectation of $y_{it}$. This allows the support of $\theta^{AS}$, $\mathcal{T}$, to be unknown. Note that the number $s$ is chosen such that the approximation error is of the same magnitude as the estimation error. The underlying key growth condition is:

$$s \log(P \vee N) = o(N) \quad (1.2.3)$$

This requires that the number covariates required to estimate $y_{it}$ is sufficiently small in comparison to the sample size.

In my consumer demand setting, this is a reasonable assumption. Consumers often will only consider items in their consideration set and they cannot consider large numbers of products every time they make a purchase decision. Further, allowing for approximate sparsity allows for us to add quadratic and other higher order terms for competing prices (Banks et al., 1997) that may or may not be included in the post-Lasso estimation.

Thus, my post-Lasso grouped fixed-effects estimator is defined as the solution of the following minimization problem:

$$\left(\widehat{\theta}^{PL}, \widehat{\alpha}^{PL}, \widehat{\gamma}^{PL}, \widehat{\mathcal{B}}^{PL}\right) = \underset{(\theta,\alpha,\gamma)\in\mathcal{B}^G\times\mathcal{A}^{GT}\times\Gamma_G}{\arg\min} \sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - x'_{it}\theta_{g_i} - \alpha_{g_i t})^2 \quad (1.2.4)$$

For given values of $\theta$ and $\alpha$, the optimal group assignment for each individual unit is:

$$\widehat{g}_i(\theta, \alpha) = \underset{g\in\{1,...,G\}}{\arg\min} \sum_{t=1}^{T}(y_{it} - x'_{it}\theta_g - \alpha_{gt})^2 \quad (1.2.5)$$

where I take the minimum $g$ in case of a non-unique solution. The standard GFE estimator relies on the usual least squares criterion function:

$$\widehat{Q}(\theta, \alpha) := \mathbb{E}_N[(y_{it} - x'_{it}\theta_{\widehat{g}_i(\theta,\alpha)} - \alpha_{\widehat{g}_i(\theta,\alpha)t})^2] \quad (1.2.6)$$

where $\widehat{g}_i(\theta, \alpha)$ is given by equation (1.2.5). The Lasso estimator is defined as a solution to the following optimization problem:

$$\left(\widehat{\theta^L}, \widehat{\alpha^L}\right) = \underset{(\theta,\alpha)\in\Theta^G\times\mathcal{A}^{GT}}{\arg\min} \widehat{Q}(\theta, \alpha) + \frac{\lambda}{n}\|\hat{\Upsilon}\theta\|_1 \quad (1.2.7)$$

where $\lambda$ is the penalty level and $\Upsilon = \text{diag}(\psi_1, ..., \psi_P)$ is a diagonal matrix specifying penalty holdings. I will use the infeasible "ideal" penalty loadings:

$$\Upsilon^0 = \text{diag}(\psi_1^0, ..., \psi_P^0) \quad \psi_j^0 = \sqrt{\mathbb{E}_n[x_{ijt}^2 v_{it}^2]} \quad j = 1, ..., P \quad (1.2.8)$$

Since $v_{it}$ is unobserved, the ideal penalty loadings are infeasible. However, by using conservative penalty loadings at first, such as $v_{it} = y_{it} - \overline{y}$, and estimating $v_{it}$ from the residuals, I can iterate to a feasible penalty loadings estimator. These penalty loadings allow for sharp convergence.

Ideal penalty loadings are used to introduce self-normalization of the first-order condition of the lasso problem. I follow the work of Belloni et al. (2012)

to use moderate deviation theory from Jing et al. (2003) to bound deviations which allows a penalty level $\lambda/N$ such that the lasso estimator converges at an oracle rate. This strategy allows the maximum of normalized scores to be well approximated by a standard normal quantile (Peña et al., 2008).

The penalty level, $\lambda/N$, should dominate the noise and this can be achieved using moderate deviation theory with the following choice for $\lambda$:

$$\lambda = c2\sqrt{n}\Phi^{-1}(1-\psi/(2P)), \tag{1.2.9}$$

$$\text{with} \quad \psi \to 0, \quad \log(1/\psi) = \mathcal{O}(\log(P \vee N))$$

I set $\psi = 0.1/\log(P \vee N)$ and $c = 1.1$ as recommended based on simulation studies in Belloni et al. (2012). This provides a theoretical strategy to use Lasso penalty that doesn't require cross-validation (which is will be very costly because of computation time) or any kind of guess work to choose our penalty value, $\lambda$. These plug-in values have been shown to be effective in estimation (Bickel et al., 2009; Drukker and Liu, 2019).

I will then use the GFE estimator applied to the model selected by Lasso. Specifically,

$$\widehat{\mathcal{T}} = \text{support}(\widehat{\theta}^L) = \{j \in \{1, ..., P\} : |\widehat{\theta}^L_{gj}| > 0\} \tag{1.2.10}$$

The set $\widehat{\mathcal{B}}$ can contain additional variables not selected by Lasso, but the number of such variables must not be larger than the number selected by Lasso. Thus, $\widehat{\mathcal{T}} \subseteq \widehat{\mathcal{B}}$. The post-Lasso estimator can be written as:

$$\left(\widehat{\theta}^{PL}, \widehat{\alpha}^{PL}\right) \in \underset{(\theta,\alpha)\in\widehat{\mathcal{B}}^G\times\mathcal{A}^{GT}}{\arg\min} \widehat{Q}(\theta, \alpha, \gamma) \tag{1.2.11}$$

where $\widehat{Q}(\theta, \alpha, \gamma)$ is defined from equation (1.2.6). Thus, my Post-Lasso estimates of $\theta$ and $\alpha$ are given by (1.2.11) and of $g_i$ is given by (1.2.5).

### 1.2.2 Computation

Equation (1.2.11) minimizes a piecewise-quadratic function, where the partition of the parameter space is defined by different values of $\widehat{g}_i(\theta, \alpha)$. However, the number of partitions of $N$ into $G$ groups increases dramatically as $N$ increases so minimizing across all partitions is in-feasible. The following algorithm can be used to minimize equation (1.2.11) and improve upon this problem.

**Algorithm 1.** *(iterative)*

1. *Let $\gamma^{(0)}$ be some starting assignment to groups. Set $s = 0$.*
2. *Compute:*

$$\left(\tilde{\theta}^{(s+1)}, \tilde{\alpha}^{(s+1)}\right) = \underset{(\theta,\alpha)\in\Theta^G\times\mathcal{A}^{GT}}{\arg\min} \mathbb{E}_n[(y_{it} - x'_{it}\theta_{g_i^{(s)}} - \alpha_{g_i^{(s)}t})] + \frac{\lambda}{n}\|\hat{\Upsilon}\theta\|_1 \tag{1.2.12}$$

3. *Select $\widehat{\mathcal{B}}^{(s+1)} \supseteq \widehat{\mathcal{T}}^{(s+1)}$ where $\widehat{\mathcal{T}}^{(s+1)}$ is defined by equation* (1.2.10)*.*

$$\left(\theta^{(s+1)}, \alpha^{(s+1)}\right) = \underset{(\theta,\alpha)\in\widehat{\mathcal{B}}^{G,(s+1)}\times\mathcal{A}^{GT}}{\arg\min} \sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - x'_{it}\theta_{g_i^{(s)}} - \alpha_{g_i^{(s)}t})^2 \tag{1.2.13}$$

4. *Compute for all $i \in \{1, ..., N\}$:*

$$g_i^{(s+1)} = \underset{g\in\{1,\ldots,G\}}{\arg\min} \sum_{t=1}^{T}(y_{it} - x'_{it}\theta_g^{(s+1)} - \alpha_{gt}^{(s+1)})^2 \tag{1.2.14}$$

5. *Set $s = s + 1$ and go to Step 2 (until numerical convergence).*

This algorithm alternates between three steps. The first step is to choose the variables that are chosen for each group using a Lasso selection technique. The second step updates $\theta$ and $\alpha$ as a post-Lasso. The last step is assigning each individual unit $i$ to the group $g_i$ which minimizes their objective function: $(y_{it} - x'_{it}\theta_g^{(s+1)} - \alpha_{gt}^{(s+1)})^2$.

Unfortunately, the solution depends on the chosen starting values since the least squares objective function is not globally convex (Bai, 2009). Thus, finding the true values requires using many starting assignments to find which one minimizes the objective function. This algorithm improves upon this and allows for computation without choosing every partition of $N$ individuals into $G$ groups. Drawing random starting values provides a practical approach for many problems.

For higher dimensional problems, I can improve on this using a process similar to the *kmeans* clustering algorithm (Forgy, 1965) which uses another algorithm that exploits recent advances in data clustering (Hansen et al., 2010). This extension, called Algorithm 2, is outlined in Appendix 1.7.1. A comparison of two very similar algoriths is done by Bonhomme and Manresa (2015). Computation remains a challenge as $N$ grows large, so there remains potential for further research to improve these techniques.

## 1.3 Asymptotic Properties

In this section, I discuss the asymptotic properties of my post-Lasso grouped fixed-effects (PL-GFE) estimator as $N$ and $T$ tend to infinity in model (1.1.1).

### 1.3.1 Setup

Consider the following data generating process:

$$y_{it} = x_{it}'\theta^0_{g_i^0} + \alpha^0_{g_i^0 t} + v_{it} \tag{1.3.1}$$

where $g_i^0 \in \{1, ..., G\}$ denotes group membership, $\theta^0_{g_i^0}$ is approximately sparse, and where the $^0$ superscripts refer to true parameter values. I assume that the number of groups $G = G^0$ is known, but I will discuss estimating the number of groups later in the paper.

Let $\left(\widetilde{\theta}, \widetilde{\alpha}\right)$ be the infeasible PL-GFE estimator where group membership $g_i$ is fixed to its population counterpart $g_i^0$:

$$\left(\widetilde{\beta}, \widetilde{\alpha}\right) = \underset{(\theta,\alpha)\in\Theta^G\times\mathcal{A}^{GT}}{\arg\min} \sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - x_{it}'\theta_{g_i^0} - \alpha_{g_i^0 t}\right)^2 . \qquad (1.3.2)$$

This is the post-Lasso estimator in the pooled regression of $y_{it}$ on $x_{it}$ and the interactions of population group dummies and time dummies.

The main result in this section provides conditions where estimated groups converge to their population counterparts and the PL-GFE estimator defined in equation (1.2.4) is asymptotically equivalent to equation (1.3.2) as $N$ and $T$ tend to infinity and $N/T^{\nu} \to 0$ for some $\nu > 0$. This allows $T$ to grow much more slowly than $N$ as long as $\nu > 1$.

### 1.3.2 Consistency

Consider the following assumptions.

**Assumption 1.1.** *There exists a constant* $M > 0$ *such that:*

*a.* $\Theta$ *and* $\mathcal{A}$ *are compact subsets of* $\mathbb{R}^P$ *and* $\mathbb{R}$*, respectively.*

*b.* $\mathbb{E}\left(||x_{it}||^2\right) \leq M$*, where* $||\cdot||$ *denotes the Euclidean norm.*

*c.* $\mathbb{E}(v_{it}) = 0$*, and* $\mathbb{E}(v_{it}^4) \leq M$*.*

*d.* $\left|\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T\sum_{s=1}^T \mathbb{E}(v_{it}v_{is}x_{it}'x_{is})\right| \leq M.$

*e.* $\frac{1}{N}\sum_{i=1}^N\sum_{j=1}^N\left|\frac{1}{T}\sum_{t=1}^T \mathbb{E}(v_{it}v_{jt})\right| \leq M.$

*f.* $\left|\frac{1}{N^2T}\sum_{i=1}^N\sum_{j=1}^N\sum_{t=1}^T\sum_{s=1}^T \mathrm{Cov}(v_{it}v_{jt}, v_{is}v_{js})\right| \leq M.$

In Assumption 1.1.a, the parameter space must be compact. Non-stationarity in the covariates and errors are ruled out in Assumptions 1.1.b and 1.1.c. These three assumptions allow us to do my analysis in a bounded space. Assumptions 1.1.d and 1.1.f impose conditions on the time series dependence of errors and covariates. Assumptions 1.1.e restricts the amount of cross-sectional dependence.

Note that this is satisfied when $v_{it}$ is independent across units. Assumptions 1.1.d-f are common in large factor models (Stock and Watson, 2002).

This assumptions must also hold for approximation error using only using $\theta$ from the support $\mathcal{T}$. Assumption 1.1.c implies that $\mathbb{E}(a(x_{it})) = 0$, and $\mathbb{E}(a(x_{it})^4) \leq M$ for some $M > 0$. This holds because the approximation error is chosen to be no larger than the model error $v_{it}$. However, I need to make further assumptions:

**Assumption 1.2.** *There exists a constant $M' > 0$ such that:*

*a.* $\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}(a(x_{it})a(x_{is})x'_{it}x_{is})\right| \leq M'$.

*b.* $\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(a(x_{it})a(x_{jt}))\right| \leq M'$.

*c.* $\left|\frac{1}{N^2T}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\operatorname{Cov}(a(x_{it})a(x_{jt}), a(x_{is})a(x_{js}))\right| \leq M'$.

Assumption 1.2 establishes the same weak dependence conditions on this error, $a(x_{it})$, as on the model error, $v_{it}$. Assumption 1.1.d. and 1.2.a. allows for lagged outcomes. In consumer demand, Assumptions 1.1 and 1.2 can allow us to use lagged prices as covariates and allow the consumers to make dynamic decisions, which can be important for some consumer demand applications (Hendel and Nevo, 2006).

In a consumer demand setting, Assumptions 1.1 restricts how much demand shocks can effect multiple individuals outside of the grouping decided by the model, as well as how much the demand shocks can last multiple periods outside of grouped demand shocks. Further, Assumption 1.2 implies that the response to other individual price changes not included in the post-Lasso estimate follow the same restrictions as the error terms. This means that these approximation error terms have restrictions over the time series dependence and how much they effect multiple individuals outside of the group effect.

Now consider the following additional assumptions on group classification:

**Assumption 1.3.**

*a. For all* $g \in \{1, ..., G\}$ : $\text{plim}_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{g_i^0 = g\} = \pi_g > 0$.

*b. There exist constants* $a > 0$ *and* $d_1 > 0$ *and a sequence* $\alpha[t] \leq e^{-at^{d_1}}$ *such that, for all* $i \in \{1, ..., N\}$ *and* $g \in \{1, ..., G\}$, $\{v_{it}\}_t$,$\{a(x_{it})\}_t$ *and* $\{\alpha_{gt}^0\}_t$ *are strong mixing processes with mixing coefficients* $\alpha[t]$*. Moreover,* $\mathbb{E}(\alpha_{gt}^0 v_{it}) = 0$ *and* $\mathbb{E}(\alpha_{gt}^0 a(x_{it})) = 0$ *for all* $g \in \{1, ..., G\}$.

*c. There exist constants* $b > 0$ *and* $d_2 > 0$ *such that* $Pr(|v_{it}| > m) \leq e^{1-(\frac{m}{b})^{d_2}}$ *for all* $i$*,* $t$*, and* $m > 0$.

*d. There exist constants* $b > 0$ *and* $d_2 > 0$ *such that* $Pr(|a(x_{it})| > m) \leq e^{1-(\frac{m}{b})^{d_2}}$ *for all* $i$*,* $t$*, and* $m > 0$.

Assumptions 1.3.a establishes that each of the $G$ population groups has a large-number of observations. Assumptions 1.3.b-d restrict the dependence and tail properties of $v_{it}$ and $a(x_{it})$. Specifically, both $v_{it}$ and $a(x_{it})$ are assumed to be strongly mixing with a faster-than-polynomial decay rate with tails also decaying at a faster-than-polynomial rate. These strengthen some aspects of Assumption 1.1. Further, $\alpha_{gt}^0$ is also assumed to be strongly mixing and contemporaneously uncorrelated with $v_{it}$ and $a(x_{it})$. This will be discussed further in Section 1.3.3. These assumptions allow me to bound misclassifation probabilities.

In the consumer demand case, these assumptions hold as long as there is the density large individual demand shocks (even if caused by price change of products left out of the model) is low. This assumption holds as long as individual demand shocks are not correlated with group aggregate demand shocks.

Now consider the following assumptions about the random coefficients model:

**Assumption 1.4.**

*a. There exists a* $\widehat{\rho} \to^p \rho > 0$ *such that, for all* $g$ : $\min_{\gamma\in\Gamma_G} \max_{\widetilde{g}\in\{1,\ldots,G\}} \widehat{\rho}(\gamma, g, \widetilde{g}) \geq \widehat{\rho}$*, where* $\widehat{\rho}(\gamma, g, \widetilde{g})$ *is the minimum* *eigenvalue of the following* $(P+T)\times(P+T)$ *matrix (with* $P = \dim x_{it}$*):*

$$M(\gamma, g, \widetilde{g}) \equiv$$

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{g_i^0 = g\}\mathbf{1}\{g_i = \tilde{g}\}\begin{pmatrix} \frac{1}{T}\sum_{t=1}^{T} x_{it}x_{it}' & \frac{1}{\sqrt{T}}x_{i1} & \frac{1}{\sqrt{T}}x_{i2} & \dots & \frac{1}{\sqrt{T}}x_{iT} \\ \frac{1}{\sqrt{T}}x_{i1} & 1 & 0 & \dots & 0 \\ \frac{1}{\sqrt{T}}x_{i2} & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{T}}x_{iT} & 0 & 0 & \dots & 1 \end{pmatrix}$$

*b. For all* $g \neq \widetilde{g}$*, there exists a* $c_{g,\widetilde{g}} > 0$ *such that* $\operatorname{plim}_{N,T\to\infty}\frac{1}{N}\sum_{i=1}^{N} D_{g\widetilde{g}i}^0 \geq c_{g,\widetilde{g}}$ *and, for all* $i \in \{1,...,N\}$*,* $\operatorname{plim}_{T\to\infty} D_{g\widetilde{g}i}^0 \geq c_{g,\widetilde{g}}$*, where* $D_{g\widetilde{g}i}^0 = \frac{1}{T}\sum_{t=1}^{T}(x_{it}'(\theta_g^0 - \theta_{\widetilde{g}}^0) + \alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0)^2$.

*c. There exists a constant* $M^* > 0$ *such that*

$$\sup_{i\in\{1,...,N\}} Pr\left(\frac{1}{T}\sum_{t=1}^{T}||x_{it}||^2 \geq M^*\right) = o(T^{-\delta})\forall\delta > 0$$

*d. For all constants* $c > 0$

$$\sup_{i\in\{1,...,N\}} Pr\left(\frac{1}{T}\sum_{t=1}^{T}||v_{it}x_{it}|| > c\right) = o(T^{-\delta})\forall\delta > 0$$

*e. For all constants* $c > 0$

$$\sup_{i\in\{1,...,N\}} Pr\left(\frac{1}{T}\sum_{t=1}^{T}||a(x_{it})x_{it}|| > c\right) = o(T^{-\delta})\forall\delta > 0$$

Assumption 1.4.a. is a relevance condition which is similar to a full rank condition in standard models. This requires that $x_{it}$ has enough within-group variation over time and across units. Bonhomme and Manresa (2015) outlines many cases where this holds. Assumption 1.4.b. is a group separation condition. Intuitively it is satisfied if, for all $i$ and $\widetilde{g} \neq g$, $\{x_{it}\}_t$ and $\{\alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0\}_t$ are not collinear. Assumption 1.4.c. holds as long as covariates have bounded support, or if they satisfy specific tail conditions. Assumption 1.4.d-e. impose further tail condition properties on $\frac{1}{T}\sum_{t=1}^{T}||v_{it}x_{it}||$ and $\frac{1}{T}\sum_{t=1}^{T}||a(x_{it})x_{it}||$. These conditions will hold if $x_{it}$ satisfies mixing and tail conditions outlined in Assumption 1.3.

In the consumer demand application, it is common to divide consumers in groups based on specific covariates like income and income level (Huang and Lin, 2007; Aasnass and Rødseth, 1983; Asano and Fiuza, 2015). However, grouping individual based on unobserved heterogeneity will allow individuals with varying covariates so I can identify the effect of these covariates within each group.

I also impose additional moment conditions on $x_{it}$ and $v_{it}$ to insure convergence of the post-Lasso coefficients.

**Condition 2.** - *The following conditions on the error terms. Let* $\widetilde{y}_{it} = y_{it} - E[y_{it}]$*:*

*a. There exists a constant* $M > 0$ *such that* $\max_{j\leq P}\mathbb{E}[\widetilde{y}_{it}^2] + \mathbb{E}[x_{itj}^2\widetilde{y}_{it}^2] + 1/\mathbb{E}[x_{itj}^2 v_{it}^2] \leq M$.

*b. There exist constants* $K_m > 0$ *such that* $\max_{j\leq P}\mathbb{E}[|x_{itj}^3 v_{it}^3|] \leq K_m$.

*c.* $K_m^2\log^3(P\vee N) = o(N)$ *and* $s\log(P\vee N) = o(N)$.

*d.* $\max_{i\leq N, t\leq T, j\leq P} x_{itj}^2[s\log(P\vee N)]/N \rightarrow^p 0$.

Parts a-b. introduce moment conditions beyond what were outlined before but follow similar intuition. Parts c-d establish growth conditions such that the number of variables needed to approximate $y_{it}$, $s$, does not grow too fast in relation to $N$.

This growth condition makes sense in my application to consumer demand. As the number of consumers increase, most of the consumers will fall into already defined groups. These groups will have the same competition sets so $s$ will not increase too quickly in relation to $N$.

To outline the next regularity condition, I must define sparse eigenvalues. To begin, I will define the $m$-sparse subset of a unit sphere as:

$$\Delta(m) = \{\delta \in \mathbb{R}^p : \|\delta\|_0 \leq m, \|\delta\|_2 = 1\},$$

and define minimal and maximal $m$-sparse eigenvalue of the gram matrix $M = \mathbb{E}_N[x_{it}x'_{it}]$ as

$$\phi_{\min}(m)(M) = \min_{\delta \in \Delta(m)} \delta' M \delta \qquad \text{and} \qquad \phi_{\max}(m)(M) = \max_{\delta \in \Delta(m)} \delta' M \delta$$

This allows me to establish the following condition:

**Condition 3.** *For any $C > 0$, there exist constants $0 < \kappa' < \kappa" < \infty$, which do not depend on $N$ but may depend on $C$, such that, with probability approaching 1, as $N, T \to \infty$,*

$$\kappa' \leq \phi_{min}(Cs)(\mathbb{E}[x_{it}x'_{it}]) \leq \phi_{max}(Cs)(\mathbb{E}[x_{it}x'_{it}]) \leq \kappa"$$

Condition 3 establishes that only certain small submatricies of the empirical Gram matrix are well-behaved. This conditions hold for standard i.i.d. sampling, but holds for more general cases as well,as outlined in Belloni and Chernozhukov (2013).

I assume that the number of variables selected by post-Lasso is not significantly larger than the model selected by Lasso. Specifically, let $\widehat{T}$ be the set of variables selected by Lasso, and $\widehat{\mathcal{B}}$ be the set of variables selected by post-Lasso. There exists some $c$ such that

$$|\widehat{\mathcal{B}} \setminus \widehat{T}| \leq c(1 \vee |\widehat{T}|) \tag{1.3.3}$$

This means that the number variables that must be included whether they are selected by the Lasso estimator or not should be relatively small. In an application to consumer demand, one can include own-price elasticity in every post-Lasso estimation but the number of cross-price elasticities or demographic effects included beyond those selected by the Lasso identification must be limited. I will not include any covariates beyond those selected by the Lasso identification in my application.

If one was focused on estimating a singular parameter such as own-price elasticity, one could use the double machine learning method (Chernozhukov et al., 2016) to solve this problem. This would significantly weaken the assumptions needed for convergence and asymptotic normality. It would also increase computation time in order to properly cross-validate the machine-learning parameters. However, my focus is on a set of parameters (cross-price elasticities as well as own-price elasticities) so I will keep my assumptions and proceed.

With these conditions, I can provide the next result which shows that my PL-GFE estimator and the infeasible least squares estimator with known population groups in equation (1.3.2) are asymptotically equivalent.

**Theorem 1.1.** *Suppose assumptions 1, 2, and 3 all hold as well as conditions 1, 2, and 3 and equation* (1.3.3)*. Also, $\lambda$ is chosen from equation* (1.2.9)*. Then, for all $\delta > 0$, as $N, T$ tend to infinity:*

$$Pr\left(\sup_{i\in\{1,\ldots,N\}} |\widehat{g}_i - g_i^0| > 0\right) = o(1) + o(NT^{-\delta}), \tag{1.3.4}$$

*and:*

$$\widehat{\theta}_g = \widetilde{\theta}_g + o_p(T^{-\delta})\forall g \tag{1.3.5}$$

*and*

$$\widehat{\alpha_{gt}} = \widetilde{\alpha_{gt}} + o_p(T^{-\delta})\forall g, t. \tag{1.3.6}$$

*Proof.* See Appendix 1.7.2. ■

The following assumptions allow to simply characterize the asymptotic distribution of the Post-Lasso estimator $(\widehat{\theta}, \widehat{\alpha})$. I denote $\overline{x}_{gt}$ as the mean of $x_{it}$ in group $g_i^0 = g$.

**Assumption 1.5.**

*a. For all* $i$*,* $j$*,* $t$*, and* $g$*:* $\mathbb{E}\left(\mathbf{1}\{g_i^0 = g\}x_{jt}v_{it}\right) = 0$.

*b. For all* $g$*, there exist positive definite matricies* $\Sigma_{\theta g}$ *and* $\Omega_{\theta g}$ *such that:*

$$\begin{aligned}\Sigma_{\theta g} &= \operatorname*{plim}_{N,T\to\infty} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{1}\{g_i^0 = g\}(x_{it}-\overline{x}_{gt})(x_{it}-\overline{x}_{gt})' \\ \Omega_{\theta g} &= \lim_{N,T\to\infty}\frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T} \\ &\qquad \mathbb{E}\left[\mathbf{1}\{g_i^0 = g\}\mathbf{1}\{g_j^0 = g\}v_{it}v_{js}(x_{it}-\overline{x}_{gt})(x_{js}-\overline{x}_{gs})'\right]\end{aligned} \tag{1.3.7}$$

*c. As* $N$ *and* $T$ *tend to infinity:* $\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{1}\{g_i^0 = g\}(x_{it}-\overline{x}_{gt})v_{it} \to^d \mathcal{N}\left(0, \sigma_{\theta g}\right)$.

*d. For all* $(g,t)$*:*
$\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}\left(\mathbf{1}\{g_i^0 = g\}\mathbf{1}\{g_j^0 = g\}v_{it}v_{jt}\right) = \omega_{gt} > 0$

*e. For all* $(g,t)$*, and as* $N$ *and* $T$ *tend to infinity:* $\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{1}\{g_i^0 = g\}v_{it} \to^d \mathcal{N}(0, \omega_{gt})$.

Assumptions 1.5.a-c. imply that the least squares estimator for $\theta_g$ has a standard asymptotic distribution. Assumption 1.5.a holds if $x_{it}$ is strictly exogenous or fixed and the observations are independent across units. In the framework of consumer demand, I will discuss this assumption in depth in Section 1.3.3. Assumptions 1.5.d-e. allow for $\alpha_{gt}$ to have a standard asymptotic distribution.

**Theorem 1.2.** *Suppose the conditions under Theorem 1.1 and Assumption 1.5 hold. Let $N$ and $T$ tend to infinity such that, for some $\nu > 0$, $N/T^{\nu} \to 0$. Then I have for all $g$:*

$$\sqrt{NT}\left(\widehat{\theta}_g - \theta_g^0\right) \to^d \mathcal{N}\left(0, [\Sigma_{\theta g}]^{-1}\Omega_{\theta g}[\Sigma_{\theta g}]^{-1}\right), \tag{1.3.8}$$

*and for all $(g,t)$:*

$$\sqrt{N}\left(\widehat{\alpha}_{gt} - \alpha_{gt}^0\right) \to^d \mathcal{N}\left(0, \frac{\omega_{gt}}{\pi_g^2}\right). \tag{1.3.9}$$

*where $\pi_g$ is defined in Assumption 3 and $\Sigma_\theta$, $\Omega_g$, and $\omega_{gt}$ are defined in Assumption 5.*

*Proof.* The proof follows the proof for Corollary 1 in Bonhomme and Manresa (2015). Note that because of our assumptions on the growth rate of $s$ and $p$ (See Condition 2), $\theta$ converges at a $\sqrt{NT}$ rate because of the penalty loadings based on the work done by Belloni et al. (2012). ■

Thus, under the conditions of Theorem 1.2, my PL-GFE estimator of $\theta^0_{g_i^0}$ is root-$NT$ consistent and asymptotically normal as long as $T$ can increase polynomially more slowly than $N$. My estimator of $\alpha^0_{g_i^0 t}$ are root-$N$ consistent and asymptotically normal. Notice that the fixed effect convergence rate is consistent with other standard time fixed-effects. Lastly, the estimated group membership indicators are uniformly consistent for the population as $N/T^{\nu} \to 0$ for some $\nu > 0$.

### 1.3.3 Price Endogeneity

In the Industrial Organization literature, there is a large concern about price endogeneity. A common issue is that prices are set in response to changes in demand, which would bias common estimates for price elasticity unless one uses instrumental variables. This becomes particularly challenging in my setting where I have a large number of prices that would need instruments. Further, in many cases estimates based on instrumental variables can be swayed by one or two data points (Young, 2019), which can be particularly dangerous with noisy demand data.

It is common to use Hausman instruments (prices in one city as an instrument in another city) to deal with price heterogeneity (Hausman et al., 1994). There can be some failure when using these instruments. For example, Hausman estimates that Kellogg Raisin Bran and Post Raisin Bran have a negative (and statistically significant) cross-price elasticity (Hausman, 1996) even though they are most likely close substitutes. Problems like this are not rare in the literature (Nevo, 2011). Many of the problems comes from the assumption about Hausman instruments that city demand shocks are uncorrelated.

Many authors have tried to avoid these problems by using other instruments, like choosing average prices of retail chains outside of the store in which the consumer is shopping (DellaVigna and Gentzkow, 2017; Hitsch et al., 2017). This instrument is chosen because retail chains change their prices over time in a coordinated manner across stores at least partially because they face a constant marginal cost (Stroebel and Vavra, 2019). They thus assume that the timing of chain-level sales is unrelated to local demand shocks. However, this does not capture how the chains may change their prices across store because of a demand shock from the group of consumers that shop at their stores.

It is worth acknowledging that the demographics of consumers is similar across stores within a chain. Thus, some have suggested to use demographics of other stores in a chain as an instrumental variable (George and Waldfogel, 2003). This has been used in estimating demand for soda and other unhealthy foods (Allcott et al., 2018, 2019). While demographics can be useful in controlling for the groups of consumers and their demand, this does not capture the unobserved heterogeneity that may dictate which stores consumers may go to.

I use the time-varying grouped fixed-effects estimator, $\alpha_{g_i t}$, and group specific slope parameters, $\theta_{g_i}$, to control for these endogeneity problems. To illustrate this, I will use a simple example of just trying to estimate one item's own-price elasticity with my model. Let $p_{it}$ be the log price of the good and $q_{it}$ be the log quantity purchased of that good.

$$q_{it} = \alpha_{g_i t} + \theta_{g_i} p_{it} + v_{it} \tag{1.3.10}$$

Based on the proof of Theorem 1.2, my estimates of $\theta_g$ will be unbiased as long as the expected value of

$$\sqrt{NT}\left(\widehat{\theta}_g - \theta_g^0\right) = \left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{1}\left\{g_i^0 = g\right\}\left(p_{it} - \overline{p}_{g_i^0 t}\right)\left(p_{it} - \overline{p}_{g_i^0 t}\right)'\right)^{-1} \cdots$$
$$\cdots\left(\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{1}\left\{g_i^0 = g\right\}\left(p_{it} - \overline{p}_{g_i^0 t}\right) v_{it}\right)$$

tends towards zero as $N$ and $T$ go to infinity where $\overline{p}_{g_i^0 t}$ is the mean of $p_{it}$ in group $g_i^0 = g$. This will hold as long as individual price deviation away from group average price is uncorrelated with individual shocks in demand. With the stores setting the prices, this does not seem like a big concern in my setting since they will respond to group changes in demand when setting prices.

Further, my estimates of $\alpha_{gt}$ will be unbiased as long as my estimates of $\theta_g$ are unbiased and

$$\sqrt{N}\left(\widehat{\alpha}_{gt} - \alpha_{gt}^0\right) = \frac{\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{1}\left\{g_i^0 = g\right\} v_{it}}{\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left\{g_i^0 = g\right\}} + o_p(1)$$

tends towards zero as $N$ goes to infinity. This will hold as long as individual group assignment is uncorrelated with individual demand shock. This should hold since individuals with correlated shocks would lead to a group time shock which is captured in the model.

I will illustrate these arguments with a few examples. On the demand side, households have a store-choice problem that effects the prices since consumers have unobserved preferences between stores and unobserved shopping costs (Allcott et al., 2017). However, these individuals can be grouped together which allows their effect to be measured by $\alpha_{g_i t}$. On the supply side, there are often price discounts around seasonal peaks in demand (Chevalier et al., 2003). Further, advertising to an individual can lead to an increase of purchases for similar consumers (Hartman, 2010). Both of these (and most other supply-side effects) would be captures by the grouped fixed effects estimator. As stores try to optimize their pricing decisions, they should be doing so by targeting groups of consumers, rather than individual consumers and my method can identify those groups as well as their demand shocks.

There is also a concern about storing products. Stores will often put their goods on sale for a week (Pesendorfer, 2002) and consumers will respond by buying that good and storing it when the good is no longer on sale (Hendel and Nevo, 2006). Estimating demand in the setting can sometimes lead to overestimates of own-price elasticity as well as underestimate some cross-price elasticities. I can minimize these problems by aggregating sales and price to a monthly level as well as including lagged variables, which I can do because of Assumptions 1.1 and 1.2.

Note that if one desires to control for endogeneity beyond the fixed effects estimators discussed in this section, a standard instrumental variable approach is not feasible using a grouped fixed-effects estimator. In order to perform statistical inference and maintain consistency of the grouped fixed-effects estimator, we must use a linear panel data. Therefore, in order to handle endogeneity, the researcher must use a control function approach similar to that used by Chernozhukov, Hausman and Newey (2019). Analysis of using this technique with my PL-GFE estimator is left for further research.

### 1.3.4 Choice of number of groups

I will follow the the analysis of large factor models (Bai and Ng, 2002) and interactive fixed-effects panel data models (Bai, 2009), to create an information criterion to consistently estimate the number of groups, $G^0$, to be used in my estimator. Consider the following class of information criteria based on model (1.1.1):

$$I(G) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - x_i'\widehat{\theta}_{\widehat{g}_i}^{(G)} - \widehat{\alpha}_{\widehat{g}_i t}^{(G)}\right)^2 + Gh_{NT}, \tag{1.3.11}$$

where $^{(G)}$ refers to the GFE estimator with $G$ groups and $h_{NT}$ is a penalty. The estimated number of groups would be defined by:

$$\widehat{G} = \underset{G\in\{1,\ldots,G_{max}\}}{\arg\min} \; I(G), \tag{1.3.12}$$

where $G_{max}$ is an upper bound on $G^0$ and is assumed to be known by the researcher[1]. Following Bai and Ng (2002) and (Bai 2009), it can be shown that the number of groups $\widehat{G}$ is consistent for $G^0$ if, as $N$ and $T$ go to infinity, $h_{nt}$ goes to zero and $(N \wedge T)h_{nt}$ tends to infinity. The first condition assures that $\widehat{G} \geq G^0$ with probability approaching one and the second guarantees that $\widehat{G} \leq G^0$.

Consider the following Bayesian Information Criterion (BIC):

$$BIC(G) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - x_i'\widehat{\theta}_{\widehat{g}_i}^{(G)} - \widehat{\alpha}_{\widehat{g}_i t}^{(G)}\right)^2 + \widehat{\sigma}^2\frac{GT+N+GP}{NT}\ln(NT), \tag{1.3.13}$$

where $\widehat{\sigma}^2$ is a consistent estimator of the variance of $v_{it}$[2]. If $N$ and $T$ go

[1] The issue of selecting $G_{max}$ is left for future research.

[2] I can use the following equation:

$$\widehat{\sigma}^2 = \frac{1}{NT - G_{max}T - N - PG_{max}}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - x_i'\widehat{\theta}_{\widehat{g}_i} - \widehat{\alpha}_{\widehat{g}_i t}\right)^2.$$

to infinity at the same rate, $\widehat{G}$ is consistent for $G^0$. If $T$ goes to infinity more slowly than $N$ such that $T/N$ tends towards zero, the BIC criterion implies $\widehat{G} \geq G^0$ but $\widehat{G}$ may be inconsistent. Some simulations that show the accuracy of the criterion are included in Appendix 1.7.3.

## 1.4 Monte Carlo Simulations

I will run simulations similar to Belloni et al. (2012). I will construct my simulations off my model:

$$y_{it} = x_{it}'\theta_{g_i} + \alpha_{g_i t} + v_{it}.$$

For simplicity, $x_{it}$, $\alpha_{g_i t}$, and $v_{it}$ are all normally distributed $N(0,1)$. I set $G = 4$. I provide results for $N = 100$ and 200, as well as $T = 7$ and 12. $\theta_{g_i}$ come from three possible distributions. There will be 20 possible covariates, $P = 20$. For the first simulation, I will set $\theta_{g_i}$ such that each group has three covariates with $\theta_{g_i} = 1$ and the rest will be set to zero, $s = 3$. There will be a similar problem where $s = 10$ which could lead the Lasso to perform poorly with the given sample size. The last simulation will use approximate sparsity rather than absolute sparsity such that for each $\theta_{g_i}$ to be a randomized order of the following sequence: $(1, 0.7, 0.7^2, ..., 0.7^{19}, 0.7^{20})$.

I will report the group fixed-effects estimator (GFE) along with my post-Lasso grouped fixed-effects estimator (PL-GFE). For each estimator, I report the median bias and median absolute deviation (MAD) for both the fixed effects, $\alpha$, and the covariate coefficients, $\theta$. I also include the rate at which units are misclassified to the wrong group (G-M). The results are reported in Table 1.1.

Simulation 1 and 2 both show that my PL-GFE out performs GFE by having a smaller bias and deviation for both $\alpha$ as well as $\theta$. It also misidentifies groups less frequently. The most dramatic improvements are for $\theta$.

For the case in Simulation 3 which is approximately sparse, there are slight improvements to the bias of $\theta$ but significant improvements to the MAD of $\theta$. These come at a sacrifice of slight increases in bias and MAD of $\alpha$. The group misidentification rate is better for PL-GFE when $N$ is small but when $N$ is sufficiently small, GFE performs just as well.

In order to prepare for analysis of consumer demand, I do these simulations again in a scenario where the covariates are correlated. Thus, $x_{it}$ is drawn from a multivariate normal distribution $\mathcal{N}(0, \Sigma)$ where $\Sigma$ is a $P \times P$ matrix that takes the value of 1 along the diagonal and 0.1 on every value off the diagonal. The value 0.1 is an approximation of the value estimated from the prices in my data. The results are shown in Table 1.2.

All three of these simulations show that when there is this much correlation between the covariates, the PL-GFE estimator always outperforms the GFE estimator. The Lasso selection accurately identifies the important covariates, while the correlation leads to bias and higher variance for the basic GFE estimator. This holds whether there is extreme sparsity, some sparsity, or approximate sparsity.

In the case of approximate sparsity, the addition of covariate covariance lead to the PL-GFE estimator significantly outperforming the GFE for the case $N = 100$ and $T = 7$. For instance, the misclassification rate drops by more than 35%. This illustrates the importance of the PL-GFE estimator in cases where $N$ and $T$ are relatively small but there are large amounts of correlated covariates.

This is helpful for my application to consumer demand since I expect prices to be correlated. Thus, using the PL-GFE will be more effective since it can determine what prices are actually influencing the purchase decisions of the consumer more accurately than the GFE would in this situation. This will hold even if the competition set is not particularly "small" or if other prices have very small effects on purchasing decisions. If price does not enter linearly, I can add higher order price terms and let the data determine which terms are important in modeling consumer behavior.

Table 1.1

Simulation Results

| | Simulation 1 (S=3) | | | | | Simulation 2 (S=10) | | | | | Simulation 3 (Exponential) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Median | | | | Mean | Median | | | | Mean | Median | | | | Mean |
| Estimator | $\alpha$-Bias | $\alpha$-MAD | $\theta$-Bias | $\theta$-MAD | G-M | $\alpha$-Bias | $\alpha$-MAD | $\theta$-Bias | $\theta$-MAD | G-M | $\alpha$-Bias | $\alpha$-MAD | $\theta$-Bias | $\theta$-MAD | G-M |
| A. $n = 100$ and $T = 7$ | | | | | | | | | | | | | | | |
| GFE | 0.1763 | 0.1765 | 0.0686 | 0.0685 | 0.0424 | 0.1743 | 0.1744 | 0.0683 | 0.0684 | 0.0379 | 0.1893 | 0.1905 | 0.0739 | 0.0738 | 0.132 |
| PL-GFE | 0.1657 | 0.1654 | 0.0097 | 0.0099 | 0.0294 | 0.1784 | 0.1846 | 0.0363 | 0.0459 | 0.0908 | 0.1939 | 0.1931 | 0.0615 | 0.0339 | 0.085 |
| B. $n = 200$ and $T = 7$ | | | | | | | | | | | | | | | |
| GFE | 0.1179 | 0.1175 | 0.0457 | 0.0457 | 0.0299 | 0.1173 | 0.1173 | 0.0459 | 0.0459 | 0.0294 | 0.1244 | 0.1243 | 0.0482 | 0.0481 | 0.0992 |
| PL-GFE | 0.1133 | 0.1132 | 0.0066 | 0.0066 | 0.0263 | 0.1155 | 0.1150 | 0.0229 | 0.0229 | 0.0265 | 0.1255 | 0.1254 | 0.0436 | 0.0232 | 0.1056 |
| C. $n = 100$ and $T = 12$ | | | | | | | | | | | | | | | |
| GFE | 0.1643 | 0.1641 | 0.0489 | 0.0488 | 0.0018 | 0.1645 | 0.1650 | 0.0491 | 0.0489 | 0.0113 | 0.1650 | 0.1653 | 0.0493 | 0.0492 | 0.0141 |
| PL-GFE | 0.1581 | 0.1580 | 0.0071 | 0.0072 | 0.0011 | 0.1608 | 0.1608 | 0.0244 | 0.0245 | 0.0024 | 0.1680 | 0.1683 | 0.0481 | 0.0234 | 0.0194 |
| D. $n = 200$ and $T = 12$ | | | | | | | | | | | | | | | |
| GFE | 0.1151 | 0.1152 | 0.0334 | 0.0333 | 0.0014 | 0.1152 | 0.1154 | 0.0334 | 0.0334 | 0.0042 | 0.1167 | 0.1167 | 0.0336 | 0.0336 | 0.0115 |
| PL-GFE | 0.1131 | 0.1132 | 0.0049 | 0.0049 | 0.0012 | 0.1147 | 0.1145 | 0.0166 | 0.0165 | 0.0037 | 0.1194 | 0.1158 | 0.0355 | 0.0167 | 0.0136 |

Results are based on 500 simulation replications and 20 possible covariates. The 20 covariates are independent. There are always 4 groups. Column labels indicate the estimation procedure. GFE is the group fixed effects estimator with heterogenous coefficients outlined by Bonhomme and Manresa (2015) while PL-GFE is the estimator outlined by this paper. I report the median bias, median absolute deviation (MAD), and group misclassification rate(G-M) for each simulation.

## Table 1.2

Simulation Results

| | Simulation 1 (S=3) | | | | | Simulation 2 (S=10) | | | | | Simulation 3 (Exponential) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Median | | | | Mean | Median | | | | Mean | Median | | | | Mean |
| Estimator | $\alpha$-Bias | $\alpha$-MAD | $\theta$-Bias | $\theta$-MAD | G-M | $\alpha$-Bias | $\alpha$-MAD | $\theta$-Bias | $\theta$-MAD | G-M | $\alpha$-Bias | $\alpha$-MAD | $\theta$-Bias | $\theta$-MAD | G-M |
| A. $n = 100$ and $T = 7$ | | | | | | | | | | | | | | | |
| GFE | 0.1769 | 0.1774 | 0.0713 | 0.0713 | 0.0507 | 0.1741 | 0.1745 | 0.0703 | 0.0702 | 0.0321 | 0.1886 | 0.1882 | 0.0763 | 0.0761 | 0.139 |
| PL-GFE | 0.1657 | 0.1656 | 0.0108 | 0.0118 | 0.0331 | 0.1691 | 0.1687 | 0.0369 | 0.0369 | 0.0296 | 0.1768 | 0.1765 | 0.0496 | 0.0527 | 0.086 |
| B. $n = 200$ and $T = 7$ | | | | | | | | | | | | | | | |
| GFE | 0.1179 | 0.1177 | 0.0476 | 0.0475 | 0.0340 | 0.1184 | 0.1184 | 0.0474 | 0.0475 | 0.0262 | 0.1253 | 0.1248 | 0.0500 | 0.0499 | 0.1001 |
| PL-GFE | 0.1142 | 0.1139 | 0.0071 | 0.0075 | 0.0305 | 0.1157 | 0.1151 | 0.0253 | 0.0267 | 0.0249 | 0.1230 | 0.1227 | 0.0381 | 0.0344 | 0.0999 |
| C. $n = 100$ and $T = 12$ | | | | | | | | | | | | | | | |
| GFE | 0.1646 | 0.1643 | 0.0506 | 0.0506 | 0.0024 | 0.1644 | 0.1641 | 0.0508 | 0.0508 | 0.0256 | 0.1652 | 0.1654 | 0.0513 | 0.0513 | 0.0164 |
| PL-GFE | 0.1587 | 0.1583 | 0.0074 | 0.0079 | 0.0021 | 0.1619 | 0.1620 | 0.0264 | 0.0277 | 0.0039 | 0.1639 | 0.1638 | 0.0411 | 0.0345 | 0.0188 |
| D. $n = 200$ and $T = 12$ | | | | | | | | | | | | | | | |
| GFE | 0.1152 | 0.1153 | 0.0349 | 0.0349 | 0.0016 | 0.1152 | 0.1151 | 0.0349 | 0.0349 | 0.0035 | 0.1166 | 0.1165 | 0.0353 | 0.0352 | 0.0124 |
| PL-GFE | 0.1132 | 0.1134 | 0.0051 | 0.0054 | 0.0014 | 0.1142 | 0.1139 | 0.0185 | 0.0194 | 0.0032 | 0.1149 | 0.1149 | 0.0292 | 0.0250 | 0.0123 |

Results are based on 500 simulation replications and 20 possible covariates. The 20 covariates are correlated. There are always 4 groups. Column labels indicate the estimation procedure. GFE is the group fixed effects estimator with heterogenous coefficients outlined by Bonhomme and Manresa (2015) while PL-GFE is the estimator outlined by this paper. I report the median bias, median absolute deviation (MAD), and group misclassification rate(G-M) for each simulation.

## 1.5 Application: Consumer Demand of Soda

I will now estimate consumer demand based on product level data using my post-Lasso grouped fixed effects estimator.

### 1.5.1 Nielsen Homescan Data

I will look at the Nielsen Scanner Data which is available through the Kilts Center at the University of Chicago Booth School of Business.[3] I will focus on the results from 2014. This is a helpful application for estimating consumption behavior since it contains detailed information based on price and quantity of purchases of many products as well as many household characteristics.

The data contains a representative sample of households in the United States that use in-home scanners to record all of their purchases intended for personal, in-home use. Nielsen matches the product scanned by the household to the actual price of the store where the product was bought. Nielsen estimates that recorded purchases account for about 30% of total household consumption. I will refer to the sum over all expenditure in the Nielsen data as total expenditure, which I will use as the relevant total outlay variable because of additive separability of the utility function. This variable can be used in derivations involving economic rationality and be the relevant "income" variable.

There are a few concerns with the data. It relies on participants successfully recording their purchases in their home, so they may suffer from some recording error. However, patterns of consumption in this data set are consistent with other commonly used data sets in the literature. Aguilar (2007) finds that life-cycle pattern of household expenditures recorded in Homescan

[3] Researcher(s) own analyses calculated (or derived) based in part on data from The Nielsen Company (US), LLC and marketing databases provided through the Nielsen Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the Nielsen data are those of the researcher(s) and do not reflect the views of Nielsen. Nielsen is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

Data is consistent with those reported for food expenditures at home in Panel Study of Income Dynamics (PSID). Einav (2010) finds that these issues are not more serious than those in any other consumption surveys like the Current Population Survey (CPS). Lin (2018) compares the fraction of expenditures on different categories of products in the Nielsen Homescan Data and finds the results consistent to results from the Consumer Expenditure Survey (CES). These issues are discussed further in Appendix 1.7.4.

I will focus on soft drink purchases. The demand for soft drinks and other comparable drinks has been examined in many settings as policymakers have been considering the impacts of “soda taxes” (See Allcott et al. (2019) for an overview. See Falbe et al. (2016), Sturm et al. (2010), and Cawley et al. (2018) for examples). In particular, there is a strong interest in soft drink consumption among different groups of individuals (Dubois et al., 2019), such as children (Han and Powell, 2013) and low-income households (Drewnowski and Specter, 2004; Currie, 2009). It is important to know what the substitution patterns are and the price elasticity for different consumers since these taxes may have a negative effect on the groups of individual they are meant to help (Allcott et al., 2018). For these reasons, I will focus on consumers that purchase large amounts of soft drinks, which is one of the major focus groups for policymakers for the US Department of Health (2016).

I will aggregate sales to a monthly level and simply estimate the demand for the most bought soft drink product (Coca-Cola) based on demographic characteristics of the individual as well as the prices of the soft drinks and other popular soft drink products based on my data. The products I examine are broken up based on brand and type of product such as 12-pack, 6-pack or size of individual pack. The exact clarification of these size are anonymized to comply with the source of the data but I will refer to them as Size A, B, C, D and E. The Coca-Cola size that I will focus on will be considered Size B for the remainder of the paper.

I only focus on consumers that purchase at least twelve units of the Coca-Cola Size B in my sample during the year and soft drinks that are sold over 20,000 times in my sample. That limits my sample to 1,721 individuals ($N = 1721$) over twelve periods ($T = 12$) and twenty other sodas besides the Coca-Cola product most purchased. My estimates for prices of sodas when consumption is zero is an average of the price faced by similar consumers shopping at the same or similar stores which is similar to a method discussed by Chernozhukov et al. (2019). A detailed description of this mechanism as well as my summary statistics for my sample are included in Appendix 1.7.4.

In this setting, I am assuming that the amount of soda each individual buys will depend on the prices of only a subset of the other sodas available. Other soda prices will have zero effect because of individual preferences and search costs. There is not enough information to identify each individual's competition set, but I will break down the sample into groups and find the competition set for each group using my PL-GFE estimator.

### 1.5.2 Coefficient Estimates

I will estimate the following model using my post-Lasso grouped-effects estimator:

$$Q_{it}^{CocaCola} = \theta_{1g_i} X_i + \theta_{2g_i} P_{it}^{CocaCola} + \theta_{3g_i} P_{it}^{OtherSoftDrinks} + \theta_{4g_i} E_{it} + \alpha_{g_i t} + v_{it} \tag{1.5.1}$$

where $Q_{it}^{CocaCola}$ is the log quantity purchased of Coca-Cola Size B products by consumer $i$ in time $t$. $X_i$ are the demographic characteristics of consumer $i$, including household income, age, education level, and number of children. $E_{it}$ is the log of household expenditure for each month to estimate total outlay ("income") elasticity of the household, $\theta_{4g_i}$. Because I do not know how many children a family has, I cannot estimate this elasticity at an individual level

## Table 1.3

Price $\theta_g$ Estimates

| | G: | 1 | 2 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Brand | Size | | | | | | | |
| Pepsi | A | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | B | 0 | $-0.289^{*}$<br>(0.151) | 0 | 0 | 0 | 0 | 0 |
| Diet P | A | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | B | 0 | 0 | 0 | $0.133^{**}$<br>(0.063) | 0 | 0 | 0 |
| MD | A | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | B | 0 | 0 | 0 | 0.076<br>(0.130) | 0 | 0 | 0 |
| Diet MD | A | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Coca-Cola | A | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | B | $-0.508^{***}$<br>(0.075) | $-0.490^{***}$<br>(0.071) | 0 | 0 | $-0.525^{***}$<br>(0.058) | $-10.522^{***}$<br>(0.238) | 0 |
| | C | 0 | 0 | -0.334<br>(0.285) | 0 | 0 | 0 | $0.527^{**}$<br>(0.217) |
| | D | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | E | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Diet CC | A | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | C | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | B | 0 | $-0.235^{**}$<br>(0.099) | 0 | 0 | 0 | 0 | $0.255^{***}$<br>(0.075) |
| CC Zero | A | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | B | 0 | 0 | 0 | 0.085<br>(0.061) | 0 | 0 | 0 |
| C-F CC Zero | A | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Dr. Pepper | A | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | B | -0.346<br>(0.212) | 0 | 0 | 0 | 0 | 0 | 0 |
| Diet Dr. P | A | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Diet P is short for Diet Pepsi. MD is short for Mountain Dew. CC is short for Coca-Cola. C-F is short for Caffeine Free. Dr. P is short for Dr. Pepper. Note that the Coca-Cola Size B is the own-price elasticities. All other estimates are cross-price elasticities. Values in parenthesis are estimates of the standard deviation. $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$

with this data. $P_{it}^{CocaCola}$ is the log price of Coca-Cola while $P_{it}^{OtherSoda}$ is the log price of the fourteen other most popular sodas for each individual $i$ in time $t$. Thus, $\theta_{2g_i}$ and $\theta_{3g_i}$ are estimates for each groups own- and cross-price elasticities respectively. I use individual clustered standard errors. I will use ten groups, which is chosen using the BIC criterion in Appendix 1.7.3.2.

I will now discuss the overall results before briefly examining the demand patterns of certain group of consumers. Table 1.3 contains my estimates for the own-price ($\theta_{2g_i}$) and cross-price elasticities ($\theta_{3g_i}$). Notice that the main competition for the Coca-Cola Size B products that I am trying to estimate is mostly other Size B products. The only other type of product whose price has a significant effect in estimating demand for Coca-Cola Size B products is Coca-Cola Size C products. Most competitors have a positive cross-price elasticity within reasonable ranges, but there are exceptions that will be discussed later. Own price elasticity generally falls around -0.5 which is similar to other estimates (Chernozhukov et al., 2019). This is because we are taking out individuals from Group 7 with high own-price elasticity. I will discuss this group further on in the paper.

Table 1.4 contains my estimates for the demographic effects on Coca-Cola Size B demand. Note that the numbers for log of overall expenditure, which can be an estimate of “income” elasticity, are positive and around 0.2, which is similar to previous estimates (see Allcott et al. (2018) for example). It may be a bit low compared to the average value because we are focused on the subgroup of consumers who purchase soda frequently. Remember that the parameters that are set to zero from the Lasso selection does not imply that the value of that coefficient is zero, but that it is close to zero and the value does not have a significant impact on predicting quantity of soda purchased. Most of the variables I originally included were not selected by the post-Lasso selection in any group so these not included in the table. Even many of the demographic variables that are included are not statistically significant. Very few demographic variables are included and statistically significant because individuals with different demographic effects are often split between different groups so there are very few significant demographic effects within each group. The demographics of each group of consumers is discussed in Appendix 1.7.5.2.

Table 1.4

Demographic $\theta_g$ Estimates

| Group: | 2 | 4 | 5 | 8 | 9 |
|---|---|---|---|---|---|
| Demographic | | | | | |
| Log(Expenditure) | 0.204$^{*}$<br>(0.119) | 0.151<br>(0.128) | 0.215$^{**}$<br>(0.104) | 0.176$^{**}$<br>(0.088) | 0.581$^{***}$<br>(0.112) |
| Male Head High School Education | -0.233<br>(0.283) | 0 | 0 | 0 | 0 |
| Married | -0.153<br>(0.278) | 0 | 0 | 0 | 0 |
| Cable | 0 | 0 | 0 | 0.180$^{*}$<br>(0.104) | 0 |
| Internet | 0 | 0 | 0.133<br>(0.261) | 0 | 0 |

Note that the first row is a measure of "income" elasticity for various groups. Further, the only demographic variables included in this table are those with at least one non-zero $\theta_g$. $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$

My estimates for group time-varying fixed effects ($\alpha_{gt}$) is displayed in Figure 1.1. The first graph contains all the group fixed effects while the second ignores two groups with fixed effects significantly different than the majority of the groups, Group 7 and Group 9. I will discuss those groups more later. Note that you cannot compare the fixed-effects across groups as differences of demand because of their different values of $\theta_g$. However, you can see how demand for Coca-Cola Size B changes over time within a group, and find group demand shocks. Additional information on the fixed values, included their fixed point estimates and their standard errors, can be found in Appendix 1.7.5.1.

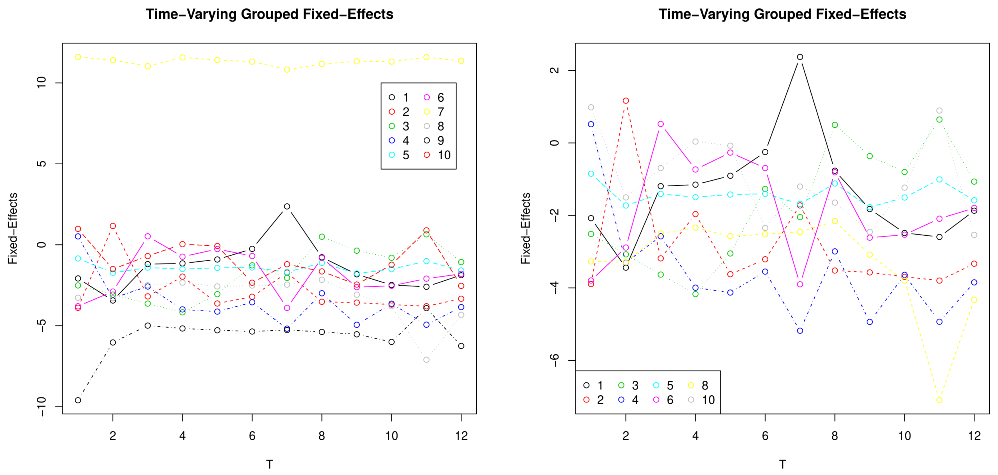

Figure 1.1: My estimates for $\alpha_{gt}$. The graph on the right excludes the case where $G = 7$ and $G = 9$.

### 1.5.3 Results Discussion

In this section, I will discuss the demand estimates of a few unique groups. A detailed examination of the demographics of each group is left for Appendix 1.7.5.2. I will reference the distribution of purchase quantities of Coca-Cola Size B products for each group, which is contained in Figure 1.3 in Appendix 1.7.5.2.

Group 7 is a particularly interesting case. The own-price elasticity is very low, while the fixed-effects are very high as well. This would imply that Coca-Cola Size B products were only purchased (or purchased at a much higher rate) when they are on sale. Further, using an indicator in the data for whether the item was on sale, I can say that the soda purchased in Group 7 was on sale 66% of the time versus just 53% of the time in non-dynamic groups. Thus, they are about 25% more likely to purchase soda on sale compared to their peers. This seems similar to the research done by (Hendel and Nevo, 2006) that consumers purchase good when they are on sale and store them during periods the goods are not on sale. Thus, for this group, an dynamic analysis would need to be done to determine the true own-price elasticity.

Group 1 has an own-price elasticity of Coca-Cola of -0.51 which is consistent with previous results. The demand of this group seems to have season trends since it increases in the summer and decreases in the winter. Of note is the negative cross-price elasticity of Dr. Pepper. This would appear counterintuitive but there are some explanations. Quantity purchased of Dr. Pepper and Coca-Cola is positively correlated and has an $R^2$ of about 0.1. Further, when Coca-Cola is purchased, the consumers are more likely to purchase 3-4 products in each time period meaning that when they buy soda, they seem to buy multiple quantities of soda. This pattern continues for groups 2 and 4. Thus, for groups 1, 2 and 4, consumers do not consume any soda most periods, but when they do, they consume more than one unit of multiple kinds of soda. Using the same type of sales analysis with group 7, I find that soda purchased by these groups are purchased on sale 64% of the time versus just 52% in non-bundling groups. These consumers appear to like bundles of soda and their demand may be more precisely estimated using a bundle choice estimation such as the methods outlined by Chintagunta and Nair (2011) or Berry et al. (2014).

Group 9 also has unique fixed-effects, but in the opposite way. Group 9 has fixed-effects that are lower than any other group each month. It also has the highest estimated income-elasticity. That seems to imply that the primary driver of how much members of this group purchase soft drinks is their income each month. It is also worth noting that members of this group also have a large negative price shock in January, that could be because of new years resolution to eat healthy or some other reason.

### 1.5.4 Implications

My PL-GFE estimator can have practical implications for firms as well as policy makers. By having group specific aggregate demand shocks, researchers can evaluate how different groups responded to certain events in time as well as group time trends. For example, if a Coca-Cola appeared in a movie that came out in July, researchers could use the PL-GFE estimator and notice that the only groups that had a positive demand shocks in that month were groups 2 and 10 while there was an actual negative shock for individuals in group 6. While these groups do not seem to be geographically concentrated in certain regions of the country, groups 2 and 10 are the two youngest groups while group 6 is the second oldest. This allows the researchers to know which types of individuals seemed to positively and negatively respond to the July event.

By having group specific price elasticities, researchers can evaluate implementation of different treatments on specific groups they are interested in. For example, when assessing the consequences of soda taxes, researchers are particularly interested in low-income households and households with children, as outlined in Section 1.5.1. I will proceed in evaluating the each of these subgroups using my PL-GFE estimates.

The group with the highest fraction of households with a yearly income lower than \$25,000 is group 8. Because this group only has positive cross-price elasticities for other Coca-Cola products, it appears that they choose Coca-Cola based on their price relative to other Coca-Cola products. This would imply that raising the price of one product would not decrease their consumption of Coca-Cola Size B products because they would substitute to another Coca-Cola product. It is also worth noting that this group had a large negative demand shock in November, so it is worth examining what happened this month to lead to this shock.

The group with the highest fraction of households that have children under 13 years of age are groups 3 and 6. Group 3 does not have any significant covariates, but it does appear to have a trend to increase their soft drink purchases throughout the year. This would imply that price did not have a significant impact on these households, so there may be better policies if the goal is to decrease soft drink consumption in these households. Group 6, however, has a significant negative own-price elasticity, which makes sense because household heads in group 6 are much less likely to have full-time jobs than household-heads in group 3.

The group that consumes the most soda overall was group 5. They are also the oldest group. This group had positive cross-price elasticities for Diet Pepsi, Mountain Dew and Coca-Cola Zero Size B products. They further have the most consistent demand across the twelve month period. They do not seem to have strong specific preference in terms of which soda they choose to consume.

These examples illustrate how researchers can use the PL-GFE estimator to provide insights that are not available through more basic or standard estimators. Researchers in firms can use it to see how demand is effected across time and before and after certain events for different groups of individuals and then further identify unique characteristics about these groups of individuals such as age and where they live. Further, policy makers can use these estimates to predict consequences of certain policies for different groups of individuals rather than solely on the aggregate or average effect of their policy.

## 1.6 Conclusion

This paper introduces post-Lasso techniques to a grouped fixed-effects model (PL-GFE) to deal with situations with a large number of variables and significant unobserved heterogeneity. I use this PL-GFE model to estimate a demand system which has many prices and heterogenous consumers with grouped time-varying demand shocks. I am able to group individuals together based on their demand shocks as well as how they respond to prices. Using grouped time-varying fixed-effects allows me to use prices in my model rather than instruments which improves precision and does not impose instrumental variable assumptions on the model.

My application was able to find significant heterogeneity among the consumers of soda. Groups had different and often unique demand shocks. We were able to identify which consumers were likely to only purchase Coca-Cola when it was on sale, as well as consumers whose Coca-Cola consumption was particularly dependent on their income. Through my estimator, we were able to examine groups with particularly high interest households (low-income households, households with children, and households that consume large quantities of soda) and understand their consumption behavior better.

There are a few ways to expand my work in this paper. One would be to include individual heterogeneity or random coefficients within each group. Combining random coefficients with lasso technique would increase the computation time which is already a major burden but it could decrease the large variation often needed to compute random-coefficients model. Although combining random coefficients with lasso has been done (Fan and Li, 2012), it has not been done with the penalty loadings (Belloni et al., 2012) which could improve them.

Further, because I make the assumption that demand shocks only effect groups of individuals for the groups as determined by the estimation procedure, individuals cannot change groups (Su et al., 2017) and the number of groups cannot change (Bonhomme et al., 2017). Thus, in a consumer demand setting I would be hesitant to extend to scenarios covering a long period of time because scenarios like this are possible. However, there may be a way to accommodate individuals changing groups or number of groups changing over time which can then accommodate settings with large $T$.

Demand estimation is often done in a discrete choice. Changing the structure of this to a logit like structure may be helpful and can be done using a Lasso-type estimator as well (Belloni et al., 2016). However, the asymptotics of the grouped fixed effects would be difficult and in its current form would introduce bias. Leaving out grouped-fixed effects would require instrumental variables if one were to try and estimate demand, which could be problematic in large-variable cases like those examined in this paper. This is left for future work.

## 1.7 Appendix

### 1.7.1 Computation

I will use the (hdm) package in R explained in (Chernozhukov et al., 2016) to accelerate the Post-Lasso computation. If I were to use Algorithm 1 in most empirical application, an infeasibly large number of starting values would need to be used to get reliable solutions. To illustrate, if I were to have $N = 1000$ and $G = 10$ then that would open $9.59 * 10^{29}$ possible combinations. Note that each starting iteration of Algorithm 1 addresses many combinations but the possible combinations is so large that the number of iterations needed to achieve the proper solution would remain infeasibly large. This can be helped by using parallel computing and it is extremely parallelizable, but there are further improvements to be made.

I can significantly decrease the computation time further by using an extension to Algorithm 1 (Bonhomme and Manresa, 2015) which introduces the Variable Neighborhood Search method (Hansen and Mladenović, 2001; Hansen et al., 2010). This would increase the number of combinations covered by each iteration of the algorithm. This specific algorithm extends from (Pacheco and Valencia, 2003) and (Brusco and Steinley, 2007). As before, let $\gamma = \{g_1, ..., g_N\}$ be a generic partition of $N$ units into $G$ groups.

**Algorithm 2.** *(Variable Neighborhood Search)*

1. *Let $\gamma_{init}$ be some starting assignment to groups.*

   *Perform steps 2-3 of Algorithm 1 to obtain Post-Lasso estimates of $(\theta, \alpha)$ based on this initial group.*

   *Set $iter_{max}$ and $neigh_{max}$ to some desired values.*

   *Set $j = 0$.*

   *Set $\gamma^* = \gamma_{init}$*

2. *Set $n = 1$.*

3. *(Neighborhood jump) Relocate $n$ randomly selected units to $n$ randomly selected groups, and obtain a new grouping $\gamma'$.*

   *Perform steps 2-3 of Algorithm 1 to obtain Post-Lasso estimates of $(\theta^{'}, \alpha^{'})$.*
4. *Set $\left(\theta^{(0)}, \alpha^{(0)}\right) = \left(\theta^{'}, \alpha^{'}\right)$ and do Algorithm 1.*
5. *(Local search) Obtain a new grouping $\gamma$" based on the $(\theta, \alpha)$ from Step 4.*
6. *If the objective function using $\gamma^{"}$ improves relative to using $\gamma^{*}$ set $\gamma^{*} = \gamma^{"}$ and go to Step 2. Otherwise, set $n = n + 1$ and go to the next step.*
7. *If $n \leq neigh_{max}$, then go to Step 3. Otherwise go to the next step.*
8. *Set $j = j + 1$. If $j > iter_{max}$, then Stop. Otherwise, go to Step 2.*

Algorithm 2 combines two different search techniques. It applies a local search (Step 5) which guarantees that a local optimum is attained since re-assigning any single individual to a different group will not decrease the objective function. Secondly, re-assigning randomly selected units into randomly selected groups (Step 3) allows for broader exploration. This is a neighborhood jump of increasing size up to size $neigh_max$ which is chosen by the researcher.

Algorithm 2 adds two parameters to be set by the researcher: the maximum neighborhood size $neigh_{max}$ and maximum number of iterations $iter_{max}$. Along with Algorithm 1, the researcher still has do determine $N_s$, the number of starting values. In my simulations, I set $N_s = 5000$, $neigh_{max} = 20$ and $iter_{max} = 20$ to guarantee convergence. The parameters set should be large enough to always yield the same result. $N_s$ can be split between processors and this algorithm is also extremely parallelizable which can make most computations feasible with this estimator.

In order to insure convergence in my empirical application with $N = 1,721$, I set $N_s = 100,000$. This is very demanding computationally, but by splitting it between 8000 nodes, it can be completed in a few days. However, once my estimator is calculated, I can use the identified groups to find trends and learn about consumers, which is not computationally intensive at all. There is a large starting cost to use my estimator, but the resulting groups and estimates have many uses that are not demanding computationally.

### 1.7.2 Proof of Theorem 1.1

*Proof.* I will follow a similar pattern to the Proposition S4 in the supplementary appendix to Bonhomme and Manresa (2015). Let $\gamma^0 = \{g_1^0, ..., g_N^0\}$ denote the population grouping. Let also $\gamma = \{g_1, ..., g_N\}$ denote any grouping of the cross-sectional units into $G$ groups. Let:

$$\widehat{\mathcal{Q}}(\theta, \alpha, \gamma) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - x'_{it}\theta_{g_i} - \alpha_{g_i t})^2 \tag{1.7.1}$$

Note that my estimator minimized $\widehat{Q}(\cdot)$ over all $(\theta, \alpha, \gamma) \in \mathcal{B}^G \times \mathcal{A}^{GT} \times \Gamma_G$. Note also:

$$\widehat{\mathcal{Q}}(\theta, \alpha, \gamma) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (\nu_{it} + a(x_{it}) + x'_{it}(\theta^{AS}_{g_i^0} - \theta_{g_i}) + \alpha^0_{g_i^{AS} t} - \alpha_{g_i t})^2 \tag{1.7.2}$$

Consider the following auxiliary objective function over the same domain.

$$\begin{aligned}\widetilde{\mathcal{Q}}(\theta, \alpha, \gamma) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (x'_{it}(\theta^{AS}_{g_i^0} - \theta_{g_i}) + \alpha^{AS}_{g_i^0 t} - \alpha_{g_i t})^2 + \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \nu_{it}^2 \\ + \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} a(x_{it})^2 + \frac{2}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \nu_{it} a(x_{it})\end{aligned} \tag{1.7.3}$$

**Lemma 1.7.1.** *Let Assumption 1.1 and Condition AS hold. Then:*

$$\underset{N,T\to\infty}{\text{plim}} \sup_{(\theta,\alpha,\gamma)\in\Theta^G\times\mathcal{A}^{GT}\times\Gamma_G} |\widehat{\mathcal{Q}}(\theta, \alpha, \gamma) - \widetilde{\mathcal{Q}}(\theta, \alpha, \gamma)| = 0 \tag{1.7.4}$$

*Proof.* I have:

$$\begin{aligned}\widehat{Q}(\theta,\alpha,\gamma)-\widetilde{Q}(\theta,\alpha,\gamma) &= \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\nu_{it}(x_{it}(\theta^0_{g_i^0}-\theta_{g_i})+\alpha^0_{g_i^0t}-\alpha_{g_it})\\ &+\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}a(x_{it})(x_{it}(\theta^0_{g_i^0}-\theta_{g_i})+\alpha^0_{g_i^0t}-\alpha_{g_it})\end{aligned}$$

I know that the first term goes to zero based on lemma S3 from the supplementary appendix of Bonhomme and Manresa (2015) based on Assumption 1.1. I will now focus on the second term.

$$\begin{aligned}&\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}a(x_{it})(x_{it}(\theta^0_{g_i^0}-\theta_{g_i})+\alpha^0_{g_i^0t}-\alpha_{g_it}) =\\ &\qquad\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}a(x_{it})x_{it}'\theta^0_{g_i^0}+\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}a(x_{it})\alpha^0_{g_i^0t}\\ &\qquad-\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}a(x_{it})\alpha_{g_it}-\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}a(x_{it})x_{it}'\theta_{g_i}\end{aligned} \tag{1.7.5}$$

Consider the last term. Note that.

$$\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}a(x_{it})x_{it}'\theta_{g_i}\right)^2 \leq \frac{1}{N}\sum_{i=1}^{N}\|\theta_{g_i}\|^2\left\|\frac{1}{T}\sum_{t=1}^{T}a(x_{it})x_{it}\right\|^2 \tag{1.7.6}$$

The left term is bounded based on Assumption 1.1.a and the right term is bounded based on Assumption 1.2.a. This holds for the first term in equation (1.7.5) as well so both of these terms are uniformly $o_p(1)$. Now I will focus on the third term in equation (1.7.5).

$$\begin{aligned}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}a(x_{it})\alpha_{g_it} &= \sum_{g=1}^{G}\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{1}\{g_i=g\}a(x_{it})\alpha_{gt}\right]\\ &=\sum_{g=1}^{G}\left[\frac{1}{T}\sum_{t=1}^{T}\alpha_{gt}\left(\sum_{i=1}^{N}\mathbf{1}\{g_i=g\}a(x_{it})\right)\right]\end{aligned}$$

Using the Cauchy-Schwartz inequality, for all $g \in \{1, ..., G\}$ :.

$$\left(\frac{1}{T}\sum_{t=1}^{T}\alpha_{gt}\left(\sum_{i=1}^{N}\mathbf{1}\{g_i = g\}a(x_{it})\right)\right)^2 \leq \left(\frac{1}{T}\sum_{t=1}^{T}\alpha_{gt}^2\right) \times \left(\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{g_i = g\}a(x_{it})\right)^2\right)$$

Assumption 1.1.a implies that the first item is uniformly bounded. I will now focus on the second item.

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{g_i = g\}a(x_{it})\right)^2 \\
&=\frac{1}{TN^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{1}\{g_i = g\}\mathbf{1}\{g_j = g\}\sum_{t=1}^{T}a(x_{it})a(x_{jt}) \\
&\leq\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\frac{1}{T}\sum_{t=1}^{T}a(x_{it})a(x_{jt})\right| \\
&=\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(a(x_{it})a(x_{jt}))\right| \\
&+\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\frac{1}{T}\sum_{t=1}^{T}(a(x_{it})a(x_{jt}) - \mathbb{E}(a(x_{it})a(x_{jt}))\right|
\end{aligned}$$

I will address the first term using Assumption 1.2.b: $\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(a(x_{it})a(x_{jt}))\right| \leq \frac{M}{N}$. Using the CS inequality on the second term gives us:

$$\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\frac{1}{T}\sum_{t=1}^{T}\right.\right. (a(x_{it})a(x_{jt}) - \mathbb{E}(a(x_{it})a(x_{jt}))|)^2 \leq \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}(a(x_{it})a(x_{jt}) - \mathbb{E}(a(x_{it})a(x_{jt}))\right)^2$$

which is bounded by $\frac{M}{T}$ based on Assumption 1.2.c. Thus, $\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}a(x_{it})\alpha_{g_i t}$ is uniformly $o_p(1)$. Thus, with the results of the inequality in (1.7.6), equation (1.7.5) is uniformly $o_p(1)$ and this ends the proof of Lemma A1. ■

Let $d_H(\theta_1, \theta_2)$ and $d_H(\alpha_1, \alpha_2)$ to denote the Hausdorff distances on $\mathbb{R}^{GP}$ and $\mathbb{R}^{GT}$, respectively, where $P = \dim x_{it}$, defined by:

$$d_H(a,b) = \max\left\{ \max_{g\in\{1,\ldots,G\}} \left( \min_{\widetilde{g}\in\{1,..,G\}} \frac{1}{T}\sum_{t=1}^{T}(a_{\widetilde{g}t} - b_{gt})^2 \right), \max_{\widetilde{g}\in\{1,..,G\}} \left( \min_{g\in\{1,\ldots,G\}} \frac{1}{T}\sum_{t=1}^{T}(a_{\widetilde{g}t} - b_{gt})^2 \right) \right\}$$

**Lemma 1.7.2.** *Let all the conditions of Theorem 1.1 hold. Then, as $N$, $T$ tend to infinity:*

$$d_H\left(\widehat{\theta}, \theta^0\right) \to^p 0, \quad \textit{and} \quad d_H\left(\widehat{\alpha}, \alpha^0\right) \to^p 0$$

*Proof.* Let $(\theta, \alpha, \gamma) \in \Theta^G \times \mathcal{A}^{GT} \times \Gamma_G$. For ease of notation, let $||\theta_g|| = (\sum_{k=1}^{P} \theta_{gk}^2)^{\frac{1}{2}}$ and $||\alpha_g|| = (\sum_{k=1}^{P} \alpha_{gt}^2)^{\frac{1}{2}}$. As in Lemma S4 of the supplementary appendix of (Bonhomme and Manresa, 2015):

$$\widetilde{\mathcal{Q}}(\theta,\alpha,\gamma) - \widetilde{\mathcal{Q}}(\theta^{AS},\alpha^0,\gamma^0) \geq \sum_{g=1}^{G} \widehat{\rho} \times \min_{\widetilde{g}\in\{1,\ldots,G\}} \left[ \left\|\theta_g^{AS} - \theta_{\widetilde{g}}\right\|^2 + \frac{1}{T}\sum_{t=1}^{T}(\alpha_{gt}^0 - \alpha_{\widetilde{g}t})^2 \right] \tag{1.7.7}$$

Where $\widehat{\rho}$ is bounded away from zero asymptotically by Assumption 1.4.a. Let

$(\widehat{\theta}, \widehat{\alpha}, \widehat{\gamma}) \in \widehat{\mathcal{B}}^G \times \mathcal{A}^{GT} \times \Gamma_G$. Remember that

$\sup_{(\theta,\alpha,\gamma)\in\Theta^G\times\mathcal{A}^{GT}\times\Gamma_G} \left[ \widetilde{\mathcal{Q}}(\widehat{\theta}, \widehat{\alpha}, \widehat{\gamma}) - \widehat{\mathcal{Q}}(\theta,\alpha,\gamma) \right] = o_p(1)$ (Bonhomme and Manresa, 2015) and $\widehat{\mathcal{B}} \subseteq \Theta$. Also note that $\widehat{\mathcal{Q}}(\widehat{\theta}, \widehat{\alpha}, \widehat{\gamma}) \leq \widehat{\mathcal{Q}}(\theta^{AS}, \alpha^0, \gamma^0)$ since I assume $\theta^{AS} \in \mathcal{T}$, from my AS assumption, and $\mathcal{T} \subseteq \widehat{\mathcal{B}}$ (Belloni and Chernozhukov, 2013). Thus, using Lemma 1.7.1:

$$\begin{aligned}\widetilde{\mathcal{Q}}(\widehat{\theta}, \widehat{\alpha}, \widehat{\gamma}) = \widehat{\mathcal{Q}}(\widehat{\theta}, \widehat{\alpha}, \widehat{\gamma}) + o_p(1) \leq \widehat{\mathcal{Q}}(\theta^{AS}, \alpha^0, \gamma^0) + o_p(1) &= \widehat{\mathcal{Q}}(\theta^0, \alpha^0, \gamma^0) + o_p(1) \\ &= \widetilde{\mathcal{Q}}(\theta^0, \alpha^0, \gamma^0) + o_p(1)\end{aligned}$$

If I combine this fact with equation (1.7.7), then I get the following:

$$\max_{g\in\{1,\dots,G\}}\left[\min_{\widetilde{g}\in\{1,\dots,G\}}\left(\left\|\theta_g^0-\widehat{\theta}_{\widetilde{g}}\right\|^2+\frac{1}{T}\sum_{t=1}^{T}(\alpha_{gt}^0-\widehat{\alpha}_{\widetilde{g}t})^2\right)\right]=o_p(1)$$

The rest of the proof follows Lemma S4 of the supplementary index of Bonhomme and Manresa (2015) and holds under Assumption 1.4.b. ■

This proof shows that there exists a permutation $\sigma:\{1,...,G\}\to\{1,...,G\}$ such that:

$$\left\|\widehat{\theta}_{\sigma(g)}-\theta_g^0\right\|^2+\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{\alpha}_{\sigma(g)t}-\alpha_{gt}^0\right)\to^p 0 \tag{1.7.8}$$

I may relabel such that $\sigma(g)=g$. For any $\eta>0$, let $\mathcal{N}_\eta$ denote the set of parameters $(\theta,\alpha)\in\Theta^G\times\mathcal{A}^{GT}$ that satisfy $\left\|\theta_g-\theta_g^0\right\|<\eta$. and $\frac{1}{T}\sum_{t=1}^{T}(\alpha_{gt}-\alpha_{gt}^0)<\eta$ for all $g\in\{1,...,G\}$. I will now work on the following result:

**Lemma 1.7.3.** *For $\eta>0$ small enough I have, for all $\delta>0$ and as $N$ and $T$ go to infinity:*

$$\sup_{(\theta,\alpha)\in\mathcal{N}_n}\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{\widehat{g}_i(\theta,\alpha)\neq g_i^0\}=o_p(T^{-\delta}).$$

*Proof.* Based on the definition of $\widehat{g}_i(\cdot)$, for all $g\in\{1,...,G\}$ :

$$\mathbf{1}\{\widehat{g}_i(\theta,\alpha)=g\}\leq\mathbf{1}\left\{\sum_{t=1}^{T}(y_{it}-x_{it}'\theta_g-\alpha_{gt})^2\leq\sum_{t=1}^{T}(y_{it}-x_{it}'\theta_{g_i^0}-\alpha_{g_i^0t})^2\right\}.$$

Now consider:

$$\begin{aligned}\mathbf{1}\{\widehat{g}_i(\theta,\alpha)\neq g_i^0\}&=\sum_{g=1}^{G}\frac{1}{N}\sum_{n=1}^{N}\mathbf{1}\{g_i^0\neq g\}\mathbf{1}\{\widehat{g}_i(\theta,\alpha)=g\}\\&\leq\sum_{g=1}^{G}\frac{1}{N}\sum_{n=1}^{N}\underbrace{\mathbf{1}\{g_i^0\neq g\}\mathbf{1}\left\{\sum_{t=1}^{T}(y_{it}-x_{it}'\theta_g-\alpha_{gt})^2\leq\sum_{t=1}^{T}(y_{it}-x_{it}'\theta_{g_i^0}-\alpha_{g_i^0t})^2\right\}}_{=Z_{ig}(\theta,\alpha)}.\end{aligned} \tag{1.7.9}$$

I will proceed by bounding $Z_{ig}(\theta, \alpha) \forall (\theta, \alpha) \in \mathcal{N}_{\eta}$, by a quantity that does not depend on $(\theta, \alpha)$. Consider:

$$
\begin{aligned}
Z_{ig}(\theta, \alpha) = \mathbf{1}\{g_i^0 \neq g\} \mathbf{1} \Bigg\{ \sum_{t=1}^{T} \Big( (\alpha_{g_i^0 t} - \alpha_g) + (x'_{it}\theta_{g_i^0} - x'_{it}\theta_g) \Big) & \, (v_{it} + a(x_{it}) \\
+ x_{it}\theta^0_{g_i^0} + \alpha^0_{g_i^0 t} - & \frac{x'_{it}\theta_g + x'_{it}\theta_{g_i^0} + \alpha_{gt} + \alpha_{g_i^0 t}}{2} \Bigg) \leq 0 \Bigg\}, \\
\leq \max_{\widetilde{g} \neq g} \mathbf{1} \Bigg\{ \sum_{t=1}^{T} \left( (\alpha_{\widetilde{g}t} - \alpha_g) + (x'_{it}\theta_{\widetilde{g}} - x'_{it}\theta_g) \right) & \, (v_{it} + a(x_{it}) \\
+ x_{it}\theta^0_{\widetilde{g}} + \alpha^0_{\widetilde{g}t} - & \frac{x'_{it}\theta_g + x'_{it}\theta_{\widetilde{g}} + \alpha_{gt} + \alpha_{\widetilde{g}t}}{2} \Bigg) \leq 0 \Bigg\},
\end{aligned}
$$

Let us define:

$$
\begin{aligned}
A_t = \Bigg| & \sum_{t=1}^{T} \left( (\alpha_{\widetilde{g}t} - \alpha_g) + (x'_{it}\theta_{\widetilde{g}} - x'_{it}\theta_g) \right) \\
& \times \left( v_{it} + a(x_{it}) + x_{it}\theta^0_{\widetilde{g}} + \alpha^0_{\widetilde{g}t} - \frac{x'_{it}\theta_g + x'_{it}\theta_{\widetilde{g}} + \alpha_{gt} + \alpha_{\widetilde{g}t}}{2} \right) \\
& - \sum_{t=1}^{T} \left( (\alpha^0_{\widetilde{g}t} - \alpha^0_g) + (x'_{it}\theta^0_{\widetilde{g}} - x'_{it}\theta^0_g) \right) \\
& \times \left( v_{it} + a(x_{it}) + x_{it}\theta^0_{\widetilde{g}} + \alpha^0_{\widetilde{g}t} - \frac{x'_{it}\theta^0_g + x'_{it}\theta^0_{\widetilde{g}} + \alpha^0_{gt} + \alpha^0_{\widetilde{g}t}}{2} \right) \Bigg|
\end{aligned}
$$

$$
\begin{aligned}
A_t \leq & \Bigg| \sum_{t=1}^{T} \left( (\alpha_{\widetilde{g}t} - \alpha_g) + (x'_{it}\theta_{\widetilde{g}} - x'_{it}\theta_g) \right) v_{it} \\
& \qquad\qquad - \sum_{t=1}^{T} \big( (\alpha^0_{\widetilde{g}t} - \alpha^0_g) + (x'_{it}\theta^0_{\widetilde{g}} - x'_{it}\theta^0_g) \big) v_{it} \Bigg| \\
& + \Bigg| \sum_{t=1}^{T} \left( (\alpha_{\widetilde{g}t} - \alpha_g) + (x'_{it}\theta_{\widetilde{g}} - x'_{it}\theta_g) \right) a(x_{it}) \\
& \qquad\qquad - \sum_{t=1}^{T} \big( (\alpha^0_{\widetilde{g}t} - \alpha^0_g) + (x'_{it}\theta^0_{\widetilde{g}} - x'_{it}\theta^0_g) \big) a(x_{it}) \Bigg| \\
& + \Bigg| \sum_{t=1}^{T} \left( (\alpha_{\widetilde{g}t} - \alpha_g) + (x'_{it}\theta_{\widetilde{g}} - x'_{it}\theta_g) \right) \left( x'_{it}\theta^0_{\widetilde{g}} - \frac{x'_{it}\theta_g + x'_{it}\theta_{\widetilde{g}}}{2} \right) - \\
& \sum_{t=1}^{T} \big( (\alpha^0_{\widetilde{g}t} - \alpha^0_g) + (x'_{it}\theta^0_{\widetilde{g}} - x'_{it}\theta^0_g) \big) \left( x'_{it}\theta^0_{\widetilde{g}} - \frac{x'_{it}\theta^0_g + x'_{it}\theta^0_{\widetilde{g}}}{2} \right) \Bigg| \\
& + \Bigg| \sum_{t=1}^{T} \left( (\alpha_{\widetilde{g}t} - \alpha_g) + (x'_{it}\theta_{\widetilde{g}} - x'_{it}\theta_g) \right) \left( \alpha^0_{\widetilde{g}t} - \frac{\alpha_{gt} + \alpha_{\widetilde{g}t}}{2} \right) - \\
& \sum_{t=1}^{T} \big( (\alpha^0_{\widetilde{g}t} - \alpha^0_g) + (x'_{it}\theta^0_{\widetilde{g}} - x'_{it}\theta^0_g) ) \big) \left( \alpha^0_{\widetilde{g}t} - \frac{\alpha^0_{gt} + \alpha^0_{\widetilde{g}t}}{2} \right) \Bigg|
\end{aligned}
$$

Using the CS inequality, for any $(\theta, \alpha) \in \mathcal{N}_\eta$:

$$
\begin{aligned}
A_T \leq & TC_1\sqrt{\eta} \left( \frac{1}{T} \sum_{t=1}^{T} \left\| v_{it}^2 \right\| \right) + TC_2\sqrt{\eta} \left( \frac{1}{T} \sum_{t=1}^{T} \left\| a(x_{it})^2 \right\| \right) \\
& + TC_3\sqrt{\eta} \left( \frac{1}{T} \sum_{t=1}^{T} \| x_{it} v_{it} \| \right) + TC_4\sqrt{\eta} \left( \frac{1}{T} \sum_{t=1}^{T} \| x_{it} a(x_{it}) \| \right) \\
& + TC_5\sqrt{\eta} \left( \frac{1}{T} \sum_{t=1}^{T} \| x_{it} \| \right) + TC_6\sqrt{\eta} \left( \frac{1}{T} \sum_{t=1}^{T} \| x_{it} \|^2 \right) + TC_7\sqrt{\eta}
\end{aligned}
$$

where $C_1$, $C_2$, $C_3$, $C_4$, and $C_5$ are constants independent of $\eta$ and $T$. Therefore:

$$
\begin{aligned}
Z_{ig}(\theta,\alpha) \leq & \max_{\widetilde{g}\neq g} \mathbf{1}\left\{ \sum_{t=1}^{T} \left( (\alpha^0_{\widetilde{g}t} - \alpha^0_{g}) + (x'_{it}\theta^0_{\widetilde{g}} - x'_{it}\theta^0_{g}) \right) ... \right. \\
& ... \left( v_{it} + a(x_{it}) + x_{it}\theta^0_{\widetilde{g}} + \alpha^0_{\widetilde{g}t} - \frac{x'_{it}\theta^0_{g} + x'_{it}\theta^0_{\widetilde{g}} + \alpha^0_{gt} + \alpha^0_{\widetilde{g}t}}{2} \right) \\
& \leq TC_1\sqrt{\eta}\left( \frac{1}{T}\sum_{t=1}^{T} \left\| v^2_{it} \right\| \right) + TC_2\sqrt{\eta}\left( \frac{1}{T}\sum_{t=1}^{T} \left\| a(x_{it})^2 \right\| \right) \\
& + TC_3\sqrt{\eta}\left( \frac{1}{T}\sum_{t=1}^{T} \left\| x_{it}v_{it} \right\| \right) + TC_4\sqrt{\eta}\left( \frac{1}{T}\sum_{t=1}^{T} \left\| x_{it}a(x_{it}) \right\| \right) \\
& \left. + TC_5\sqrt{\eta}\left( \frac{1}{T}\sum_{t=1}^{T} \left\| x_{it} \right\| \right) + TC_6\sqrt{\eta}\left( \frac{1}{T}\sum_{t=1}^{T} \left\| x_{it} \right\|^2 \right) + TC_7\sqrt{\eta} \right\}
\end{aligned}
$$

The right hand side of the inequality in the indicator function does not depend on $(\theta, \alpha)$, so define: $sup_{(\theta,\alpha)\in\mathcal{N}_\eta} Z_{ig}(\theta,\alpha) \leq \widetilde{Z}_{ig}$, where:

$$
\begin{aligned}
\widetilde{Z}_{ig}(\theta,\alpha) = & \max_{\widetilde{g}\neq g} \mathbf{1}\left\{ \sum_{t=1}^{T} \left( (\alpha^0_{\widetilde{g}t} - \alpha^0_{g}) + (x'_{it}\theta^0_{\widetilde{g}} - x'_{it}\theta^0_{g}) \right) (v_{it} + a(x_{it})) \leq \right. \\
& - \frac{1}{2}\sum_{t=1}^{T} \left( (\alpha^0_{\widetilde{g}t} - \alpha^0_{g}) + (x'_{it}\theta^0_{\widetilde{g}} - x'_{it}\theta^0_{g}) \right)^2 + TC_1\sqrt{\eta}\left( \frac{1}{T}\sum_{t=1}^{T} \left\| v^2_{it} \right\| \right) \\
& + TC_2\sqrt{\eta}\left( \frac{1}{T}\sum_{t=1}^{T} \left\| a(x_{it})^2 \right\| \right) + TC_3\sqrt{\eta}\left( \frac{1}{T}\sum_{t=1}^{T} \left\| x_{it}v_{it} \right\| \right) \\
& + TC_4\sqrt{\eta}\left( \frac{1}{T}\sum_{t=1}^{T} \left\| x_{it}a(x_{it}) \right\| \right) + TC_5\sqrt{\eta}\left( \frac{1}{T}\sum_{t=1}^{T} \left\| x_{it} \right\| \right) \\
& \left. + TC_6\sqrt{\eta}\left( \frac{1}{T}\sum_{t=1}^{T} \left\| x_{it} \right\|^2 \right) + TC_7\sqrt{\eta} \right\}
\end{aligned}
$$

Thus, by bringing this back to equation (1.7.9),

$$
\sup_{(\theta,\alpha)\in\mathcal{N}_\eta} \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\widehat{g}_i(\theta,\alpha) \neq g^0_i\} \leq \frac{1}{N}\sum_{i=1}^{N}\sum_{g=1}^{G} \widetilde{Z}_{ig}. \tag{1.7.10}
$$

Take $\widetilde{M} > \max(\sqrt{M}, M^*)$, where $M$ and $M^*$ are given by assumptions 1.1 and 1.4.c, respectively. Note that $\mathbb{E}(v^2_{it}) \leq \sqrt{M}$ and $\mathbb{E}(a(x_{it})^2) \leq \sqrt{M}$ because the approximation error has the same magnitude as the model error.Now consider:

$$
\begin{aligned}
Pr(\widetilde{Z}_{ig}=1) \leq & \sum_{\widetilde{g}\neq g} Pr\left(\sum_{t=1}^{T}\left((\alpha^0_{\widetilde{g}t}-\alpha^0_g)+(x'_{it}\theta^0_{\widetilde{g}}-x'_{it}\theta^0_g)\right)(v_{it}+a(x_{it}))\right.\\
& \leq -\frac{1}{2}\sum_{t=1}^{T}\left((\alpha^0_{\widetilde{g}t}-\alpha^0_g)+(x'_{it}\theta^0_{\widetilde{g}}-x'_{it}\theta^0_g)\right)^2+TC_1\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}\left\|v^2_{it}\right\|\right)\\
& + TC_2\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}\left\|a(x_{it})^2\right\|\right)+TC_3\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}v_{it}\|\right)\\
& + TC_4\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}a(x_{it})\|\right)+TC_5\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|\right)\\
& + TC_6\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|^2\right)+TC_7\sqrt{\eta}\\
\leq & \sum_{\widetilde{g}\neq g}\left[Pr\left(\frac{1}{T}\sum_{t=1}^{T}\left\|v^2_{it}\right\|\geq\widetilde{M}\right)+Pr\left(\frac{1}{T}\sum_{t=1}^{T}\left\|a(x_{it})^2\right\|\geq\widetilde{M}\right)\right.\\
& + Pr\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}v_{it}\|\geq C_1\right)+Pr\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}a(x_{it})\|\geq C_2\right)\\
& + Pr\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|\geq\widetilde{M}\right)+Pr\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|^2\geq\widetilde{M}\right)\\
& + Pr\left(\frac{1}{T}\sum_{t=1}^{T}\left((\alpha^0_{\widetilde{g}t}-\alpha^0_g)+(x'_{it}\theta^0_{\widetilde{g}}-x'_{it}\theta^0_g)\right)^2\leq\frac{c_{g,\widetilde{g}}}{2}\right)\\
& + Pr\left(\sum_{t=1}^{T}\left((\alpha^0_{\widetilde{g}t}-\alpha^0_g)+(x'_{it}\theta^0_{\widetilde{g}}-x'_{it}\theta^0_g)\right)(v_{it}+a(x_{it}))\right.\\
& \qquad \leq -T\frac{c_{g,\widetilde{g}}}{4}+T(C_1+C_2)\sqrt{\eta}\sqrt{\widetilde{M}}\\
& \qquad\quad \left.\left.+T(C_3+C_4+C_7)\sqrt{\eta}+T(C_5+C_6)\sqrt{\eta}\widetilde{M}\right)\right]
\end{aligned}
\tag{1.7.11}
$$

For the rest of the proof, I will take advantage of Lemma B5 from Bonhomme and Manresa (2015).

**Lemma.** ***B5*** *Let $z_t$ be a strongly mixing process with zero mean, with strong mixing coefficients $\alpha[t] \leq e^{-at^{d_1}}$ and with tail probabilities $Pr(|z_t| > z) \leq e^{1-(\frac{z}{b})^{d_2}}$, where $a$, $b$, $d_1$ and $d_2$ are positive constants. Then, for all $z > 0$ I* *have, for all $\delta > 0$:*

$$T^{\delta} Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T} z_t\right| \geq z\right) \rightarrow^{T\rightarrow\infty} 0.$$

By assumptions 1.1.a. and 1.4.b, I have $\lim_{T\rightarrow\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left((\alpha^0_{\widetilde{g}t}-\alpha^0_g)+(x'_{it}\theta^0_{\widetilde{g}}-x'_{it}\theta^0_g)\right)^2\right] = c_{g,\widetilde{g}}$.Thus, for $T$ large enough:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left((\alpha^0_{\widetilde{g}t}-\alpha^0_g)+(x'_{it}\theta^0_{\widetilde{g}}-x'_{it}\theta^0_g)\right)^2\right] \geq \frac{2c_{g,\widetilde{g}}}{3}$$

Now I can apply Lemma B5 to $z_t = \left((\alpha^0_{\widetilde{g}t}-\alpha^0_g)+(x'_{it}\theta^0_{\widetilde{g}}-x'_{it}\theta^0_g)\right)^2 - \mathbb{E}\left[\left((\alpha^0_{\widetilde{g}t}-\alpha^0_g)+(x'_{it}\theta^0_{\widetilde{g}}-x'_{it}\theta^0_g)\right)^2\right]$ which satisfies appropriate mixing and tail conditions because of Assumptions 1.1.a. and 1.3.c.. Take $z = \frac{c_{g,\widetilde{g}}}{6}$ and for all $\delta > 0$ and as $T$ tends to infinity:

$$Pr\left(\frac{1}{T}\sum_{t=1}^{T}\left((\alpha^0_{\widetilde{g}t}-\alpha^0_g)+(x'_{it}\theta^0_{\widetilde{g}}-x'_{it}\theta^0_g)\right)^2 \leq \frac{c_{g,\widetilde{g}}}{2}\right) = o\left(T^{-\delta}\right) \qquad (1.7.12)$$

Continuing with this reasoning, let $z_t = v^2_{it} - \mathbb{E}\left[v^2_{it}\right]$ and $z = \widetilde{M} - \sqrt{M}$ and using Lemma B5 gets us:

$$Pr\left(\frac{1}{T}\sum_{t=1}^{T} v^2_{it} \geq \widetilde{M}\right) = o\left(T^{-\delta}\right) \qquad (1.7.13)$$

for all $\delta > 0$. This holds because $\{v^2_{it}\}_t$ is strongly mixing since $\{v_{it}\}_t$ is strongly mixing from Assumption 1.3.b. The same logic gets us the following as well:

$$Pr\left(\frac{1}{T}\sum_{t=1}^{T} a(x_{it})^2 \geq \widetilde{M}\right) = o\left(T^{-\delta}\right) \qquad (1.7.14)$$

For the next step, let $c$ be the minimum of $c_{g,\widetilde{g}}$ over all $g \neq \widetilde{g}$.

$$\eta \leq \left(\frac{c}{8\left((C_1+C_2)\sqrt{\widetilde{M}}+(C_3+C_4+C_7)+(C_5+C_6)\widetilde{M}\right)}\right)^2 \qquad (1.7.15)$$

This provides an upper bound on $\eta$, so using an $\eta$ that satisfies this upper bound, I get:

$$\begin{aligned}
&Pr\left(\sum_{t=1}^{T}\left((\alpha^0_{\widetilde{g}t}-\alpha^0_g)+(x'_{it}\theta^0_{\widetilde{g}}-x'_{it}\theta^0_g)\right)(v_{it}+a(x_{it}))\right.\\
&\qquad\qquad\left.\le -T\frac{c_{g,\widetilde{g}}}{4}+T(C_1+C_2)\sqrt{\eta}\widetilde{M}^{3/2}+T(C_3+C_4)\sqrt{\eta}\widetilde{M}+TC_5\sqrt{\eta}\right)\\
&\quad\le Pr\left(\left((\alpha^0_{\widetilde{g}t}-\alpha^0_g)+(x'_{it}\theta^0_{\widetilde{g}}-x'_{it}\theta^0_g)\right)(v_{it}+a(x_{it}))\le -\frac{c_{g,\widetilde{g}}}{8}\right)
\end{aligned}$$

Because of Assumptions 1.3.b-d, $\left\{\left((\alpha^0_{\widetilde{g}t}-\alpha^0_g)+(x'_{it}\theta^0_{\widetilde{g}}-x'_{it}\theta^0_g)\right)(v_{it}+a(x_{it}))\right\}_t$ has zero mean, is strongly mixing with faster than polynomial decay rate and satisfies the tail condition of Lemma B5. Thus, let $z_t=\left((\alpha^0_{\widetilde{g}t}-\alpha^0_g)+(x'_{it}\theta^0_{\widetilde{g}}-x'_{it}\theta^0_g)\right)(v_{it}+a(x_{it}))$ and $z=\frac{c_{g,\widetilde{g}}}{8}$.

$$Pr\left(\left((\alpha^0_{\widetilde{g}t}-\alpha^0_g)+(x'_{it}\theta^0_{\widetilde{g}}-x'_{it}\theta^0_g)\right)(v_{it}+a(x_{it}))\le -\frac{c_{g,\widetilde{g}}}{8}\right)=o\left(T^{-\delta}\right) \quad (1.7.16)$$

Plugging in equations (1.7.12), (1.7.13), (1.7.14), and (1.7.16) into equation (1.7.11) and using Assumptions 1.4.c-e. I get:

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\sum_{g=1}^{G}Pr\left(\widetilde{Z}_{ig}=1\right)\le&G(G-1)\sup_{i\in\{1,\dots,N\}}\left[Pr\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}v_{it}\|\ge C_1\right)\right.\\
&+Pr\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}a(x_{it})\|\ge C_2\right)\\
&+Pr\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|\ge \widetilde{M}\right)\\
&\left.+Pr\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|^2\ge \widetilde{M}\right)\right]+o(T^{-\delta})\\
=&o(T^{-\delta})
\end{aligned}$$

To complete the proof for Lemma 1.7.3: Choose $\eta$ that satisfies equation

(1.7.15). For all $\delta > 0$ and $\epsilon > 0$.

$$Pr\left(\sup_{(\theta,\alpha)\in\mathcal{N}_\eta}\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{\widehat{g}_i(\theta,\alpha)\neq g_i^0\} > \epsilon T^{-\delta}\right) \leq Pr\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{g=1}^{G}\widetilde{Z}_{ig} > \epsilon T^{-\delta}\right) \leq \frac{\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{g=1}^{G}\widetilde{Z}_{ig}\right)}{\epsilon T^{-\delta}} = o(1) \tag{1.7.17}$$

■

The rest of the proof of this theorem follows Theorem 1.2 of Bonhomme and Manresa (2015). ■

### 1.7.3 Number of Groups

I will simulate to estimate the capability of the BIC estimator in Equation (1.3.13).Then I will use it to estimate the number of groups in my estimation application.

#### 1.7.3.1 Groups Simulations

I will use each of the simulations from Section 1.4. Remember that the first case is very sparse with three non-zero coefficients and 17 zero coeffients; the second case is sparse with ten non-zero and ten zero coefficients; the last case is approximately sparse. For each of the simulations, I simulate four scenarios, when $G = 4$ and when $G = 10$ and when $N = 100$ and $N = 200$. The simulation is run 500 times.

The simulation results from Table 1.5 compare quite favorably compared to those done by Bonhomme and Manresa (2015). This makes sense because their simulations rely on covariate coefficients being consistent across groups, which increases the overall variation significantly compared to my model where coefficients are different for each group of individuals.

As long as the results are sufficiently sparse, such as in Simulation 1, then the BIC criterion applied to my PL-GFE estimator performs very well and correctly identifies the number of groups over 99% of the time in each case of $N$ and $G^0$ simulated.

There are significant errors when $N$ is small relative to $G$ and the sparsity assumption may not be fully satisfied. Specifically, the correct number of groups in Simulation 2, when $S = 10$, is misidentified more than 30% of the time when $N = 100$ and $G^0 = 10$. However, if $N$ increases to 200, the misidentification almost ceases completely. A similar trend exists in Simulation 3, but it is less magnified. The overall trend that if the sparsity assumption may not be satisfied, a large $N$ will be required to properly identify $\widehat{G}$.

#### 1.7.3.2 Groups for Consumer Demand of Soft Drinks

Here I will apply the methodology outlined by Section 1.3.4 to my soft drink data outlined in Section 1.5 to choose the number of groups used in my estimation. The calculated Baisian Information Criterion is reported in Table 1.6.

Thus, I will use $G = 10$ or ten groups because it has the lowest BIC estimates. $N$ is sufficiently large with respect to $G$ based on my simulations such that I feel confident that it is the best choice to explain the consumer behavior in my data.

All of the numerical values of these criteria are pretty similar and this is not the only way to choose the number of groups. One can choose the number of groups based on prior beliefs of what the groups will be (Bonhomme and Manresa, 2015). For instance, in a consumer demand case, you may be interested in which consumers only buy products when they are on sale and stores products while they are not on sale, such as in group 7 in my study. These consumers can be estimated dynamically using the model from Hendel and Nevo (2006).

Table 1.5

Choice of Number of Groups (BIC)

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Simulation 1<br>S=3 | | | | | | | | | | | |
| | $N = 100$ | | | | | | $N = 200$ | | | | |
| | $G^0 = 4$ | | | | | | $G^0 = 4$ | | | | |
| $G =$ | 2 | 3 | 4 | 5 | 6 | $G =$ | 2 | 3 | 4 | 5 | 6 |
| $\%(\widehat{G} = G)$ | 0 | 0 | 100 | 0 | 0 | $\%(\widehat{G} = G)$ | 0 | 0 | 99.8 | 0.2 | 0 |
| | $G^0 = 10$ | | | | | | $G^0 = 10$ | | | | |
| $G =$ | 8 | 9 | 10 | 11 | 12 | $G =$ | 8 | 9 | 10 | 11 | 12 |
| $\%(\widehat{G} = G)$ | 0 | 0 | 99.6 | 0.2 | 0.2 | $\%(\widehat{G} = G)$ | 0 | 0 | 100 | 0 | 0 |
| Simulation 2<br>S=10 | | | | | | | | | | | |
| | $N = 100$ | | | | | | $N = 200$ | | | | |
| | $G^0 = 4$ | | | | | | $G^0 = 4$ | | | | |
| $G =$ | 2 | 3 | 4 | 5 | 6 | $G =$ | 2 | 3 | 4 | 5 | 6 |
| $\%(\widehat{G} = G)$ | 0 | 5.6 | 93.2 | 1.2 | 0 | $\%(\widehat{G} = G)$ | 0 | 0 | 99.2 | 0.8 | 0 |
| | $G^0 = 10$ | | | | | | $G^0 = 10$ | | | | |
| $G =$ | 8 | 9 | 10 | 11 | 12 | $G =$ | 8 | 9 | 10 | 11 | 12 |
| $\%(\widehat{G} = G)$ | 22 | 7 | 67.6 | 3 | 0.4 | $\%(\widehat{G} = G)$ | 0 | 0.2 | 99.8 | 0 | 0 |
| Simulation 3<br>Exponential | | | | | | | | | | | |
| | $N = 100$ | | | | | | $N = 200$ | | | | |
| | $G^0 = 4$ | | | | | | $G^0 = 4$ | | | | |
| $G =$ | 2 | 3 | 4 | 5 | 6 | $G =$ | 2 | 3 | 4 | 5 | 6 |
| $\%(\widehat{G} = G)$ | 0 | 0.2 | 97.6 | 2.2 | 0 | $\%(\widehat{G} = G)$ | 0 | 0 | 98.8 | 1.2 | 0 |
| | $G^0 = 10$ | | | | | | $G^0 = 10$ | | | | |
| $G =$ | 8 | 9 | 10 | 11 | 12 | $G =$ | 8 | 9 | 10 | 11 | 12 |
| $\%(\widehat{G} = G)$ | 5.2 | 2.2 | 90.4 | 2.0 | 0.2 | $\%(\widehat{G} = G)$ | 0.2 | 0.4 | 99 | 0.4 | 0 |

Table 1.6

Demand Estimation BIC Estimates

| G = | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|
| BIC = | 13.62 | 13.38 | 13.23 | 13.12 | 13.05 | 12.98 | 12.93 | 12.97 | 12.98 |

This table reports the Bayesian information criterion for the number of groups, $G$. $G_{max}$ is set at 15.

By choosing two groups, you may be able to identify which consumers purchase goods dynamically versus statically.

Choosing groups may sometimes be more art than science. However, the purpose of my PL-GFE estimator is to eliminate research discretion as much as possible. Researchers can include as many variables as they would like so they do not need to choose specific variables that they think will matter. The PL-GFE estimator will determine which variables are important to which individuals based on the data and not researcher imputed parameters. For this reason, I recommend using BIC to allow the data to determine the number of groups.

### 1.7.4 Nielsen Scanner Data

There are a few issues to keep in mind when dealing with this Homescan data. The first issue is with misreporting of quantity. Einav et al. (2010) examines which goods are more likely to be subject to this error. They find that consumable goods like small drinks (like many soft drinks) is likely to be consumed before getting home so are more likely to not be scanned. There are also recording errors such as when a six-pack of goods are purchased and recorded as quantity six. These are both problems that can add noise to my results, but should not bias my results because quantity is only a dependent variable in my model.

Another source of measurement error that is more concerning can come from the price. Individuals record their purchases by scanning the items they buy when they get home. The individuals input the quantity they purchase and

Nielsen matches it with the average price of the good at the store where they purchased it that week. This can lead to two types of errors. The first comes from the price changing in the middle of the week. These types of errors are approximately normally distributed.

The second type of error comes from not including discounts from loyalty cards. Einav et al. (2010) examines a retailer used in the Homescan data which has loyalty cards and finds that loyalty cards are used in about 75-80% of the transactions. Further, this would bias my prices upwards, which when comparing Homescan data with data from the retailer finds that the prices used in the Homescan data is about 7% higher. On the other hand, these price measurement errors may be overestimated since some retailers do not have loyalty cards at all. Further, Homescan data errors are comparable to errors found in other commonly used data sets (Einav et al., 2010; Aguiar and Hurst, 2007; Lin, 2018). Additional examination of this measurement error and it's effect on the results is left for future research.

When there is no good purchased, I attempt to find the average price for each month at the store the most commonly purchase soft drinks at by matching with Nielsen Retail Scanner Data. If I am unable to identify the store where the individual commonly purchases soft drinks in the month, of if the store's prices are unavailable, I estimate the prices based on average prices paid by similar consumers. The subset of similar consumers I choose is explained below. If there were no prices in the subset I tried to match, I moved to a broader subset below.

1. Individuals with the same favorite retail chain, income level, county and zip-code
2. Individuals with the same favorite retail chain, income level, and county
3. Individuals with the same favorite retail chain, and zip-code
4. Individuals with the same favorite retail chain, and county

5. Individuals with the same favorite retail chain, income level, and designated market
6. Individuals with the same favorite retail chain, and designated market
7. Individuals with the same zip-code
8. Individuals with the same county
9. Individuals with the same designated market
10. Individuals with the same favorite retail chain
11. All individuals

One could also calculate prices based on an average price consumer paid in different time periods or by using price they paid in the period before or after. These different methods do not change the results significantly and my method allows more price variation over time for each individual. Summary statistics for the quantity and price of each product is included in Table 1.7 Remember that I focused my estimator on the group of individuals who averaged one purchase of Coca-Cola Size B products in a month. This group of individuals is not representative of the populations, and the demographics of the group is included in Table 1.8.

### 1.7.5 Consumer Demand Estimation Results

Below are additional results from my estimation on the demand for soda from Section 1.5. First, I will examine the estimates for the time-varying grouped fixed-effects ($\alpha_{gt}$). Then I will examine the different demographics of the different groups of consumers.

#### 1.7.5.1 Fixed Effects

Table 1.9 contains the means and standard deviations of each groups time-varying fixed-effects, $\alpha_{gt}$. Remember that comparing different fixed-effects across groups is infeasible because of different $\theta_g$ values. For instance, Group 7

Table 1.7

| Brand | Size | Price | Quantity |
|---|---|---|---|
| Pepsi | A | 1.373 | 0.069 |
| | | (0.530) | (0.740) |
| | B | 3.702 | 0.212 |
| | | (0.846) | (1.114) |
| Diet Pepsi | A | 1.590 | 0.024 |
| | | (1.739) | (0.454) |
| | B | 2.995 | 0.040 |
| | | (1.814) | (0.456) |
| Mountain Dew | A | 1.363 | 0.043 |
| | | (1.826) | (0.444) |
| | B | 3.76 | 0.126 |
| | | (0.745) | (0.775) |
| Diet Mountain Dew | A | 1.673 | 0.014 |
| | | (1.739) | (0.358) |
| Coca-Cola | A | 1.428 | 0.441 |
| | | (0.353) | (2.280) |
| | B | 3.671 | 2.340 |
| | | (1.067) | (3.301) |
| | C | 3.004 | 0.236 |
| | | (0.634) | (2.613) |
| | D | 1.66 | 0.165 |
| | | (0.234) | (1.136) |
| | E | 1.173 | 0.112 |
| | | (1.289) | (2.228) |
| Diet Coca-Cola | A | 1.538 | 0.059 |
| | | (1.837) | (0.816) |
| | B | 3.060 | 0.196 |
| | | (1.651) | (1.194) |
| | C | 2.692 | 0.023 |
| | | (1.903) | (0.385) |
| Coca-Cola Zero | A | 1.668 | 0.042 |
| | | (1.936) | (0.755) |
| | B | 2.993 | 0.077 |
| | | (1.727) | (0.652) |
| Caffeine Free Coca-Cola Zero | A | 1.69 | 0.005 |
| | | (1.875) | (0.145) |
| Dr. Pepper | A | 1.39 | 0.045 |
| | | (0.796) | (0.508) |
| | B | 3.79 | 0.173 |
| | | (1.651) | (0.953) |
| Diet Dr. Pepper | A | 1.678 | 0.009 |
| | | (1.767) | (0.211) |

Price average is considered its median average monthly price and quantity average is its mean monthly quantity for each individual. Standard deviation is included in the parenthesis.

Table 1.8

| Demographic | Value | Household | Male | Female |
|---|---|---|---|---|
| Household Size | 2 | 0.424 | | |
| | $\geq 3$ | 0.437 | | |
| Income Level | $<$ 25,000 | 0.150 | | |
| | $> 50,000$ & $< 100,000$ | 0.381 | | |
| | $> 100,000$ | 0.131 | | |
| Age | $< 35$ | | 0.210 | 0.145 |
| | $> 55$ | | 0.436 | 0.442 |
| No Head | | | 0.179 | 0.088 |
| Type of Residence | Multi-Family house | 0.099 | | |
| | Mobile-home | 0.060 | | |
| Household Composition | Live with spouse | 0.704 | | |
| | Live with roommates | 0.020 | | |
| | Live with family | 0.135 | | |
| Children | Teenager | 0.136 | | |
| | Child | 0.170 | | |
| Job Status | Full-time | | 0.468 | 0.316 |
| | Part-time | | 0.083 | 0.167 |
| Education | High School | | 0.768 | 0.876 |
| | College | | 0.255 | 0.293 |
| Marriage Status | Married | 0.729 | | |
| | Previously married | 0.178 | | |
| Race | Black/African-American | .051 | | |
| | Asian | .023 | | |
| | Hispanic | .061 | | |
| Region | New England | 0.044 | | |
| | Mid-Atlantic | 0.083 | | |
| | East North Central | 0.219 | | |
| | West North Central | 0.080 | | |
| | South Atlantic | 0.207 | | |
| | East South Central | 0.087 | | |
| | West South Central | 0.136 | | |
| | Mountain | 0.068 | | |
| Appliances | Owns listed appliances | 0.034 | | |
| Cable | Subscription | 0.769 | | |
| Internet | Connection | 0.951 | | |

has the largest fixed-effects for every time period, but individuals in that group bought less Coca-Cola on average than individuals outside of that group.

Table 1.9

Fixed-Effects ($\alpha_{gt}$) Estimates

| Group: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Month | | | | | | | | | | |
| January | -2.08 | -3.89 | -2.51 | 0.52 | -0.85 | -3.80 | 11.61 | -3.27 | -9.60 | 0.98 |
| | (1.10) | (1.19) | (1.73) | (1.37) | (1.03) | (1.20) | (1.16) | (1.06) | (1.17) | (1.29) |
| February | -3.44 | 1.17 | -3.07 | -3.30 | -1.73 | -2.89 | 11.40 | -3.29 | -6.03 | -1.50 |
| | (1.09) | (1.18) | (1.75) | (1.40) | (1.02) | (1.18) | (1.15) | (1.08) | (1.19) | (1.31) |
| March | -1.19 | -3.19 | -3.61 | -2.58 | -1.41 | 0.52 | 11.03 | -2.50 | -4.99 | -0.69 |
| | (1.11) | (1.18) | (1.74) | (1.39) | (1.05) | (1.20) | (1.14) | (1.07) | (1.18) | (1.32) |
| April | -1.15 | -1.97 | -4.17 | -3.99 | -1.50 | -0.73 | 11.57 | -2.34 | -5.17 | 0.04 |
| | (1.12) | (1.18) | (1.73) | (1.37) | (1.05) | (1.19) | (1.16) | (1.04) | (1.14) | (1.29) |
| May | -0.91 | -3.62 | -3.05 | -4.13 | -1.42 | -0.27 | 11.41 | -2.57 | -5.28 | -0.08 |
| | (1.14) | (1.18) | (1.78) | (1.33) | (1.03) | (1.18) | (1.13) | (1.07) | (1.18) | (1.31) |
| June | -0.25 | -3.21 | -1.27 | -3.55 | -1.40 | -0.69 | 11.32 | -2.52 | -5.36 | -2.34 |
| | (1.14) | (1.18) | (1.74) | (1.38) | (1.04) | (1.21) | (1.15) | (1.08) | (1.19) | (1.30) |
| July | 2.37 | -1.72 | -2.05 | -5.18 | -1.68 | -3.90 | 10.82 | -2.46 | -5.26 | -1.20 |
| | (1.09) | (1.20) | (1.73) | (1.38) | (1.01) | (1.19) | (1.14) | (1.07) | (1.15) | (1.31) |
| August | -0.77 | -3.52 | 0.50 | -2.99 | -1.12 | -0.81 | 11.17 | -2.16 | -5.38 | -1.65 |
| | (1.13) | (1.18) | (1.74) | (1.37) | (1.03) | (1.20) | (1.15) | (1.06) | (1.17) | (1.35) |
| September | -1.82 | -3.57 | -0.36 | -4.94 | -1.77 | -2.61 | 11.33 | -3.08 | -5.53 | -2.45 |
| | (1.11) | (1.17) | (1.72) | (1.36) | (1.05) | (1.22) | (1.15) | (1.05) | (1.15) | (1.32) |
| October | -2.48 | -3.69 | -0.80 | -3.64 | -1.51 | -2.53 | 11.32 | -3.80 | -6.00 | -1.23 |
| | (1.01) | (1.19) | (1.74) | (1.38) | (1.04) | (1.19) | (1.16) | (1.08) | (1.16) | (1.32) |
| November | -2.59 | -3.80 | 0.64 | -4.93 | -1.01 | -2.09 | 11.59 | -7.10 | -3.91 | 0.89 |
| | (1.11) | (1.19) | (1.74) | (1.37) | (1.03) | (1.20) | (1.13) | (1.04) | (1.17) | (1.30) |
| December | -1.87 | -3.33 | -1.07 | -3.85 | -1.58 | -1.80 | 11.36 | -4.33 | -6.25 | -2.54 |
| | (1.12) | (1.20) | (1.78) | (1.38) | (1.03) | (1.19) | (1.17) | (1.08) | (1.19) | (1.30) |

The estimates for each $\alpha_{gt}$ is listed above with the standard deviation of each estimate listed below in parenthesis.

However, one can compare fixed-effects over time within each group. For this purpose, I include graphs of each groups' time-varying fixed-effects along with their 95% confidence bands in Figure 1.2. There you can see some groups do not change demand over time (see Groups 5 and 7), some groups general increasing or decreasing trends across multiple time periods (see Groups 1 and 3), and some groups have single time period shocks (see Groups 2 for a positive shock and Group 9 for a negative shock).

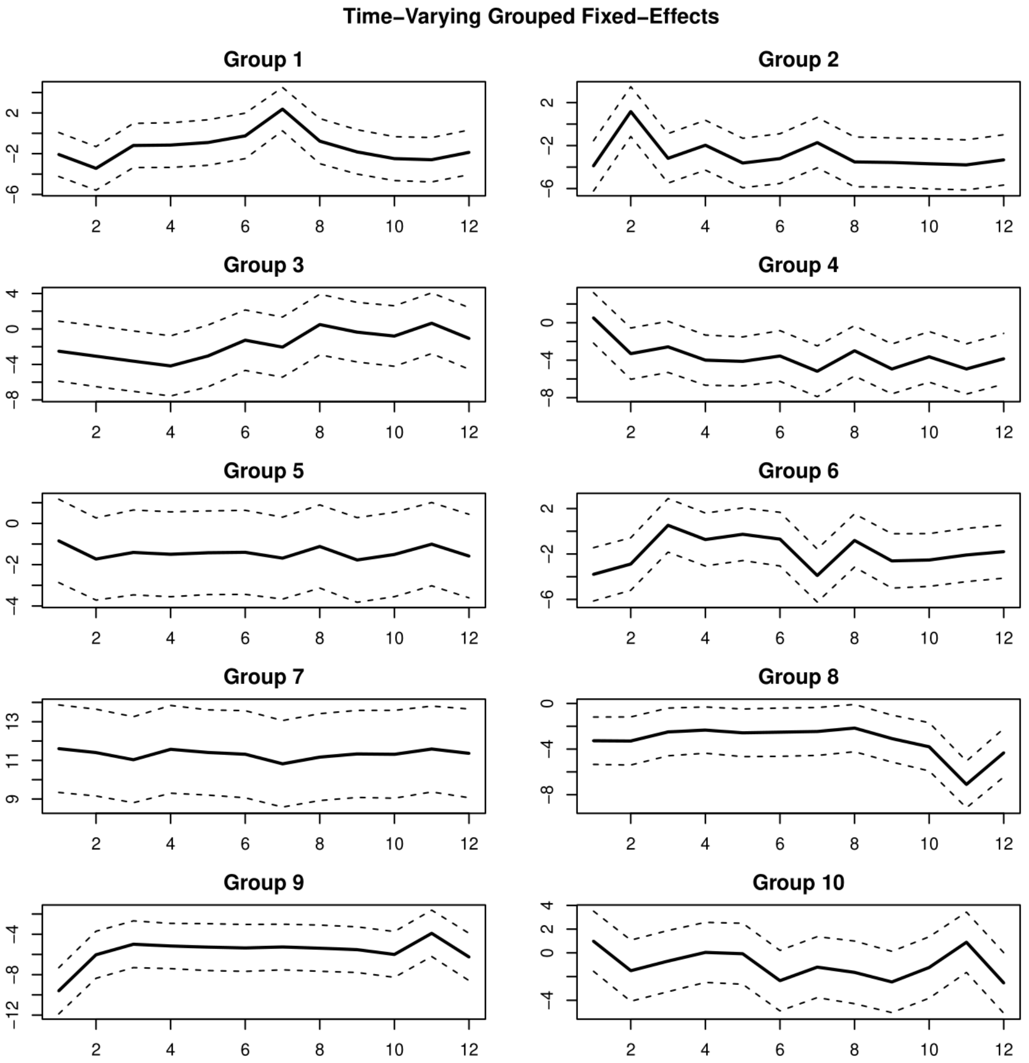

Figure 1.2: The solid lines are my PL-GFE estimates the the dotted lines are the 95% confidence bands of my estimates.

#### 1.7.5.2 Demographics of Groups

Since I know which individuals are in each group, I can estimate the demographics of each group and compare the groups with each other. The estimates for the average for the demographic variables I included in my model is shown in Table 1.10. Remember that the PL-GFE estimator groups individuals based on their response to prices, $\theta_g$, and their time-varying fixed-effects, $\alpha_{gt}$. Thus, demographics does not directly factor into groupings so the groupings in my case do not have distinct demographic characteristics, but there are important differences that could tell us more about the groups.

Further, with knowledge of group membership I know the purchase history of each individual in the group. This can allow me to see which products are more often bought in one group or another. I will use this information to graph the distribution of Coca-Cola Size B purchases of each group in Figure 1.3. With these resources and the estimates from my PL-GFE estimator, I will expound on information I know about each of the groups of consumers. This discussion expands on the discussion in Section 1.5.3.

Group 1, which has Dr. Pepper as a complement to Coca-Cola, consumes less soda overall than any other group. As discussed previously, they don't purchase Coca-Cola often, but when they do they normally purchase more than one unit of it and often with Dr. Pepper. They appear to consume more Coca-Cola in the Summer and less in the Winter. Compared to other groups they are more likely to live along the East Coast of the United States in states like Massachusetts, Connecticut, Maryland and Virginia, while being less likely to live in the middle of the United States in states like Utah, Colorado, Ohio and Michigan. Compared to most other groups, there are high-income families with a stay-at home mother.

Group 2, which has Pepsi and Diet Coca-Cola as complements for Coca-Cola, has lower total soda expenditure than every other group besides Group

## Table 1.10.A

Group Summary Statistics

| Group: | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Demographic | | | | | |
| Log(Expenditure) | 8.656 | 8.604 | 8.612 | 8.532 | 8.803 |
| | (0.580) | (0.577) | (0.561) | (0.578) | (0.596) |
| Soda Expenditure | 102.93 | 110.26 | 140.28 | 116.42 | 288.20 |
| | (67.28) | (80.14) | (122.68) | (83.47) | (178.72) |
| CC Bought | 18.07 | 18.61 | 22.12 | 17.43 | 60.20 |
| | (6.62) | (6.33) | (11.15) | (5.33) | (29.38) |
| Household Size = 2 | 0.405 | 0.386 | 0.393 | 0.436 | 0.442 |
| Household Size $\geq 3$ | 0.458 | 0.503 | 0.441 | 0.443 | 0.401 |
| Income $< 25{,}000$ | 0.137 | 0.157 | 0.159 | 0.164 | 0.143 |
| $> 50,000$ & $< 100,000$ | 0.353 | 0.451 | 0.393 | 0.379 | 0.387 |
| $> 100,000$ | 0.157 | 0.111 | 0.124 | 0.121 | 0.120 |
| Age $< 35$ | 0.216 | 0.222 | 0.241 | 0.221 | 0.194 |
| Age $> 55$ | 0.399 | 0.444 | 0.414 | 0.436 | 0.498 |
| Teenager Dependent | 0.144 | 0.144 | 0.083 | 0.107 | 0.106 |
| Child Dependent | 0.196 | 0.203 | 0.228 | 0.136 | 0.092 |
| Male Full Time Job | 0.477 | 0.444 | 0.524 | 0.486 | 0.415 |
| Female Full Time Job | 0.209 | 0.288 | 0.310 | 0.357 | 0.327 |
| High School | 0.758 | 0.732 | 0.759 | 0.757 | 0.797 |
| College | 0.248 | 0.268 | 0.255 | 0.229 | 0.253 |
| New England | 0.078 | 0.039 | 0.062 | 0.043 | 0.041 |
| Mid-Atlantic | 0.078 | 0.118 | 0.076 | 0.093 | 0.101 |
| East North Central | 0.150 | 0.183 | 0.193 | 0.264 | 0.235 |
| West North Central | 0.059 | 0.098 | 0.083 | 0.057 | 0.051 |
| South Atlantic | 0.288 | 0.176 | 0.214 | 0.157 | 0.244 |
| East South Central | 0.098 | 0.072 | 0.076 | 0.121 | 0.101 |
| West South Central | 0.144 | 0.163 | 0.193 | 0.143 | 0.120 |
| Mountain | 0.026 | 0.085 | 0.034 | 0.036 | 0.060 |
| Sample Size | 153 | 153 | 145 | 140 | 217 |

These are the means for each of these demographic variables for each group. The continuous variables includes a standard deviation in parenthesis as well. Log(Expenditure) is the log of the monthly expenditure of the household. Soda Expenditure is the total spent on soda by the household in the year. CC Bought is the number of Coca-Cola Size B products bought by each household. If not specified, the variable is applied to the male head of the household. Sample size is the number of individuals in each group. For each demographic variable, the highest value is highlighted in green while the lowest is highlighted in red.

# Table 1.10.B

## Group Summary Statistics

| Group: | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|
| Demographic | | | | | |
| Log(Expenditure) | 8.640 | 8.590 | 8.703 | 8.660 | 8.626 |
| | (0.578) | (0.518) | (0.588) | (0.617) | (0.583) |
| Soda Expenditure | 124.29 | 127.59 | 172.81 | 170.54 | 143.47 |
| | (87.38) | (122.63) | (201.16) | (123.44) | (84.11) |
| CC Bought | 18.96 | 24.80 | 30.25 | 31.07 | 27.19 |
| | (8.44) | (15.43) | (17.13) | (14.75) | (13.85) |
| Household Size = 2 | 0.363 | 0.429 | 0.457 | 0.461 | 0.449 |
| Household Size $\geq 3$ | 0.462 | 0.432 | 0.431 | 0.404 | 0.411 |
| Income $<$ 25,000 | 0.163 | 0.094 | 0.181 | 0.177 | 0.165 |
| $> 50,000$ & $< 100,000$ | 0.263 | 0.424 | 0.394 | 0.369 | 0.361 |
| $> 100,000$ | 0.150 | 0.173 | 0.106 | 0.106 | 0.120 |
| Age $< 35$ | 0.206 | 0.191 | 0.234 | 0.213 | 0.184 |
| Age $> 55$ | 0.463 | 0.451 | 0.410 | 0.440 | 0.373 |
| Teenager Dependent | 0.150 | 0.150 | 0.164 | 0.199 | 0.108 |
| Child Dependent | 0.225 | 0.162 | 0.181 | 0.156 | 0.152 |
| Male Full Time Job | 0.406 | 0.552 | 0.457 | 0.454 | 0.430 |
| Female Full Time Job | 0.293 | 0.387 | 0.309 | 0.277 | 0.348 |
| High School | 0.744 | 0.786 | 0.771 | 0.745 | 0.804 |
| College | 0.213 | 0.289 | 0.255 | 0.284 | 0.234 |
| New England | 0.025 | 0.045 | 0.027 | 0.021 | 0.057 |
| Mid-Atlantic | 0.081 | 0.060 | 0.080 | 0.078 | 0.076 |
| East North Central | 0.188 | 0.252 | 0.213 | 0.227 | 0.266 |
| West North Central | 0.088 | 0.090 | 0.085 | 0.113 | 0.076 |
| South Atlantic | 0.213 | 0.177 | 0.197 | 0.199 | 0.209 |
| East South Central | 0.144 | 0.056 | 0.101 | 0.057 | 0.057 |
| West South Central | 0.081 | 0.094 | 0.181 | 0.149 | 0.127 |
| Mountain | 0.088 | 0.105 | 0.074 | 0.078 | 0.063 |
| Sample Size | 160 | 266 | 188 | 141 | 158 |

These are the means for each of these demographic variables for each group. The continuous variables includes a standard deviation in parenthesis as well. Log(Expenditure) is the log of the monthly expenditure of the household. Soda Expenditure is the total spent on soda by the household in the year. CC Bought is the number of Coca-Cola Size B products bought by each household. If not specified, the variable is applied to the male head of the household. Sample size is the number of individuals in each group. For each demographic variable, the highest value is highlighted in green while the lowest is highlighted in red.

1. Like Group 1, they don't purchase Coca-Cola often, but when they do they normally purchase more than one unit of it and often with other sodas. They also appear to have a demand shock in February. Compared to households in other groups, households in Group 2 have more than two individuals in the household, are most likely to have annual income between \$50,000 and \$100,000. The male heads of these households are least likely to have a High-School education. This appears to be working class households.

Group 4, which has Coca-Cola Size C as a complement for Coca-Cola Size B, but the least Coca-Cola Size B overall and make the fewest overall purchases compared to every other group. Similar groups 1 and 2, they do not frequently buy Coca-Cola Size B products but when they do, they purchase large quantities of Coca-Cola Size B products and often do so along with Coca-Cola Size C products. They had a demand shock of Coca-Cola in January. There are no other demographic characteristics unique to this group, but it is the smallest out of all of the groups.

The Group 3 PL-GFE estimate selected no covariates in the model, so the estimation comes from its estimates of $\alpha_{gt}$. There is a trend in the $\alpha_{gt}$ such that it increases over time, implying that Coca-Cola 12-pack consumption increases throughout the year. Compared to other groups, households in group 3 are often younger and are more likely to have young children in the household. Group 9 is similar in that $\alpha_{gt}$ is lowest in January and increases to a steady level after March. Thus, in both of these groups, consumers that consume less in the beginning of the year, perhaps because of a new-years resolution, and increase throughout the year. Group 9 also has the highest income-elasticity out of all of the groups. Group 3 are much more likely to have young children in the home, while Group 9 is much more likely to have a teenager in the home.

The Group 10 PL-GFE estimate selected no covariates in the model, so the estimation comes from its estimates of $\alpha_{gt}$. There seems to be a general trend in the $\alpha_{gt}$ such that it decreases over time besides a demand shock in November. Compared to other groups, the male head of household is most likely to be between 35 and 55 and they are the most likely to have graduated from high school. It may be worth what is driving these households to their unique demand time-trend.

Group 5 consumes the most goods overall, soda overall and Coca-Cola compared to other groups. They have no distinct time trend in their fixed-effects but have substitutes in Diet Pepsi, Mountain Dew and Coca-Cola Zero. Compared to other groups, these are often older households without children living at home.

Group 8 has a negative demand shock in November and has substitutes in Coca-Cola Size C products, and Diet Coca-Cola Size B products. They have the lowest income compared to individuals in other groups, but purchase soda and Coca-Cola more than any group besides Group 5. However, they are more likely to have income below $50,000 and least likely to have income over $100,000. They are also the most likely to have children living in the household. These appear to be low-income households that purchase Coca-Cola consistently and will choose the cheapest product from their competition set of Coca-Cola products.

The most unique characteristic of group 6 is the shape of its fixed effects graph, since it appears to have multiple demand shocks. Compared to other groups, this group is most likely to have a household size of one and least likely to have a male with a full time job in the household. They are also least likely to have gone to college. This group has many unique demographic characteristics and multiple unique demand shocks. This group might be individuals that respond to advertising or other cultural trends, but I cannot tell what with the data I have.

Group 7 is the group that appears to follow the consumer model of Hendel

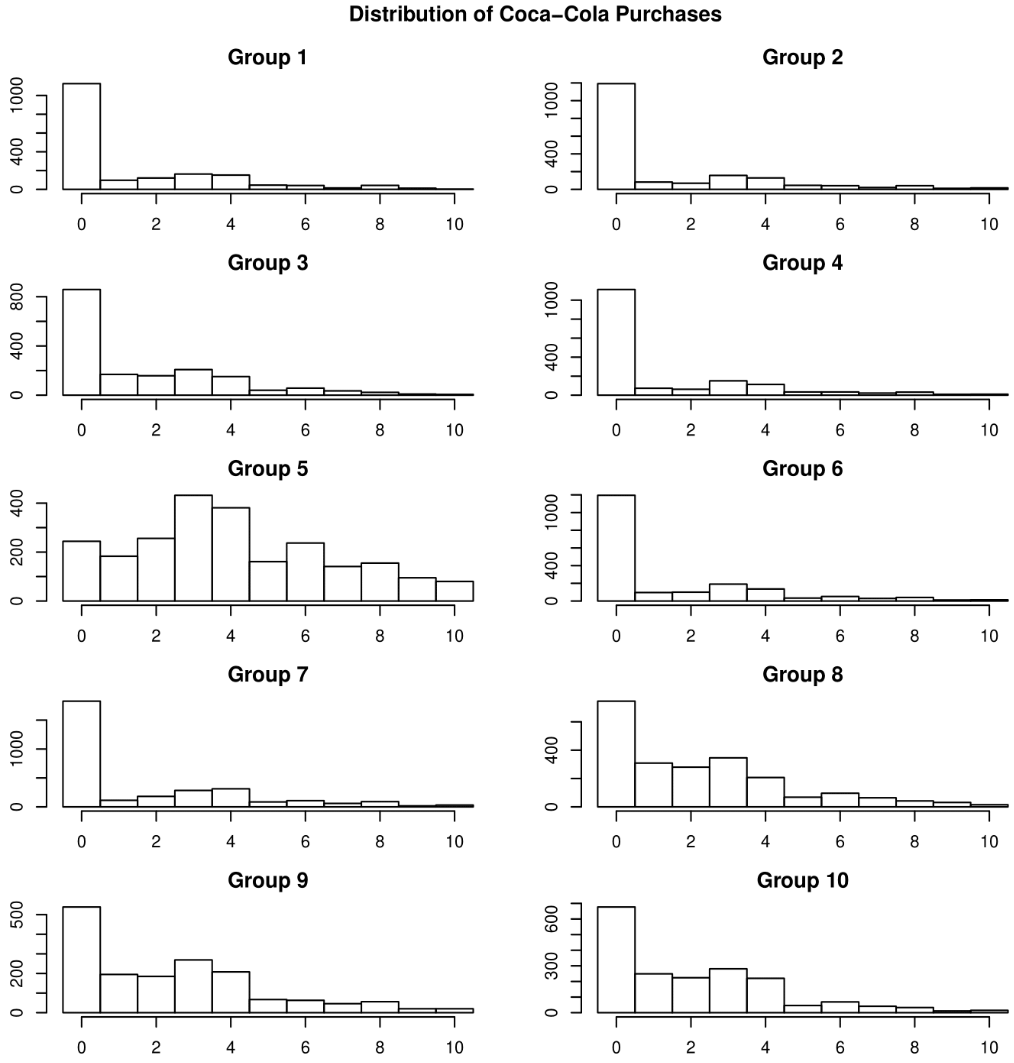

Figure 1.3: These show the distribution of the amount of Coca-Cola Size B products that are purchased each month for each group.

and Nevo (2006). There doesn't appear to be a demand shock for individuals at specific times. These individuals appear to have higher income, more likely to have a full-time job and more likely to have graduated from college compared to other groups. Also, compared to other groups they are less likely to live in the Mid-Atlantic, which is the most densely populated region which would make sense because storage would be more expensive for these households. This is also the largest group.

# Chapter 2

# A Panel Data Estimator for the Distribution and Quantiles of Marginal Effects in Nonlinear Structural Models with an Application to the Demand for Junk Food

## 2.1 Introduction

It is commonplace that panel data allows researchers to model the impact of correlated unobserved individual specific heterogeneity, as is illustrated by the fixed effects approach and generalizations to linear random coefficients models (Chamberlain, 1982; Wooldridge, 2005; Graham and Powell, 2012; Arellano and Bonhomme, 2012). A particular challenge, however, arises with the presence of nonlinearities in many microeconometric models, even in models that do not feature a limited dependent variable. This situation arises frequently in economics. While economic models often exhibit qualitative restrictions stemming from constrained optimization of rational agents, e.g., convexity or monotonicity, they feature linearity or additivity only in exceptional cases. In consumer demand which motivates the application of this paper, this has led to the rise and popularity of nonlinear models (e.g., the QUAIDS, (Banks et al., 1997)), and nonparametric and nonseparable models in general, because they capture important aspects of the data that are otherwise missed.

But while it is now commonly found that microeconomic relationships should allow for nonlinearities on the individual level, there is even more experimental and observational evidence that individuals differ across the population in ways that are not entirely captured by observable variables. There are basically two ways to deal with this complex unobserved heterogeneity: considering average effects, or recovering the distribution of heterogeneity parameters. The former is easier to obtain than the latter, and frequently less stringent assumptions have to be imposed for its recovery. As a case in point, in a cross-section setup, average treatment effects are identified under general conditions, while to recover heterogeneous functions or parameters one has to, for instance, impose monotonicity of the structural function in a scalar unobservable (Matzkin, 2003), or a linear random coefficients structure (Hoderlein and Mammen, 2007). Moreover, when covariates are endogenous, further restrictions are necessary (Imbens and Newey, 2009; Kasy, 2011; Hoderlein et al., 2017).

This paper establishes the strength of panel data to allow recovery of the distribution of heterogeneous nonparametric marginal effects, even if covariates are correlated and the time span considered is very short. More precisely, we show that the distribution of marginal effects of a general class of structural models is nonparametrically identified. This allows for arbitrary dependence between the time-invariant unobservable and the covariates of interest, provided as little as two observations are available for the individuals. Formally, we consider a nonparametric and heterogeneous model of the form

$$Y_{k,t} \;=\; \Phi(X_{k,t}, A_k) + U_{k,t}\,, \qquad k = 1, \ldots, n;\;, t = 1, \ldots, T\,, \tag{2.1.1}$$

where $Y_{k,t} \in \mathcal{Y} \subseteq \mathbb{R}$, and $X_{k,t} \in \mathcal{X} \subseteq \mathbb{R}^J$ are observable variables, and $A_k \in \mathcal{A} \subseteq \mathbb{R}^\infty$ and $U_{k,t} \in \mathcal{U} \subseteq \mathbb{R}$ are unobserved. Note that in this model, the dimension of $A_k$ is not restricted, and the structural function $\phi$ is assumed to be smooth in the sense of being twice continuously differentiable in $x_j$ for all $j = 1, .., J$, with bounded second derivatives, but is otherwise unrestricted. Moreover, we allow for arbitrary dependence (correlation) between any element of $A_k$ and any element of $X_{k,t}$ for any $k, t$. These facts make our model different from the models of Altonji and Matzkin (2005) and Evdokimov (2010) with which it shares structural similarities.

The main result in this paper establishes nonparametric identification of the (marginal) distribution of marginal effects $\partial_{x_j}\phi(x, A)$, for $j = 1, \ldots, J$, and all $x \in \mathcal{X}$, even with many regressors and only two time periods (i.e., $T = 2$). If $T \geq J + 1$, we also show that the joint distribution of all marginal effects, i.e., $\nabla_x\phi(x, A) = (\partial_{x_1}\phi(x, A), \ldots, \partial_{x_J}\phi(x, A))'$ is identified, for all $x \in \mathcal{X}$, see Remark 2.4. As a corollary, we obtain identification of objects like the average structural marginal effect, as well as the variance of marginal effects. An important limitation of our analysis is that we can only make statements for the population for which $X_{k,1} = X_{k,2} = ... = X_{k,T}$, i.e., we are only identifying the distribution $f_{\nabla_x\phi(x,A)|X_1-X_2=0}$ for the "stayers" (in the sense of Chamberlain (1982)). To fix ideas, in our demand application this will be the population for which income and prices stay approximately constant. As an important contribution, we establish that this limitation is not an accident of the identification approach taken, but a consequence of a profound non-identification result for nonlinear marginal effects outside of the stayers sub-population. The intuition behind this result is as follows: Suppose the true model is a $J$-th order polynomial in a scalar $X_{it}$ with random coefficients on every term. Then, the number of time periods acts as limiting factor for our ability to learn about this complex models - if $J$ exceeds $T - 1$,; there is generic non-identification (i.e., with $T = 2$, at most a linear random coefficients model is identified for $x_2 \neq x_1$).

The essential idea which underlies this strong constructive identification result for the stayers is as follows: Unlike with repeated cross section data, we utilize the fact that we observe individuals repeatedly in a panel to form a derivative dependent variable $\partial Y/\partial X$. Specifically, by considering individuals whose $X_{k,2}$ is close to their $X_{k,1}$ we construct a sample counterpart to the limiting process when taking derivatives. A complication arises because we have to correct for the transitory error $U_{k,t}$. This is done by considering people who have exactly $X_{k,2} = X_{k,1} = x$ for every $x \in \mathcal{X}$ (or, in the sample, almost exactly), because for these individual all changes in $Y_{k,t}$ can be attributed to changes in $U_{k,t}$. In the sample, we thus use the difference between people who are at or very near the diagonal from those who are near, but not quite as near, to the diagonal. The difference in the distribution of $Y_{k,t}$ is then due to the (heterogeneous) causal marginal effect of $X_{k,t}$. This effect depends, obviously, in general on the position $X_{k,2} = X_{k,1} = x$ we consider; by letting the position $x$ vary, we obtain an arbitrary nonlinear relationship. Fig. 1 illustrates the population used in the sample. Finally, that this works only near the diagonal (i.e., only for the stayers) is due to the fact that higher order terms in the derivative approximation only disappear in this neighborhood.

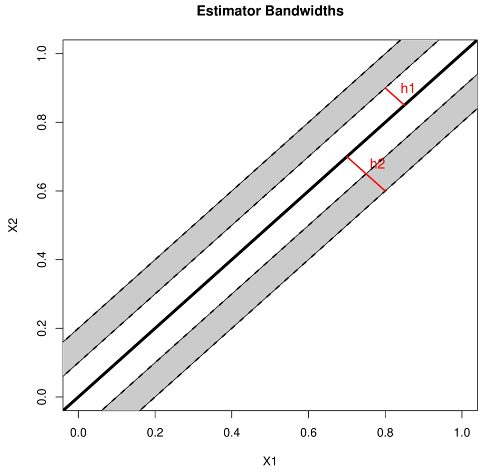

Figure 2.1: The shaded region is the region we will use, which is more than $h_1$ away from where $X_1 = X_2$ but less than $h_2$ away from where $X_1 = X_2$.

The baseline specification allows us to identify the marginal distribution of every marginal effect needing only two time periods. However, its driving force is the time invariance of the unobservable as well as the structural function. With more time periods, we may relax this assumption and allow for the structural relationship to change over time under restrictions on the way time enters which may be weakened as $T$ becomes large. Several other extensions are briefly discussed in this paper: The approach can be augmented to allow for discrete covariates; however, the effect of interest has to be on a continuous variable. More generally, we may control for additional covariates through a semiparametric specification. Finally, we conjecture that the approach can be extended to a discrete dependent variable if one exogenous regressor with large support is available, similar to Honoré and Lewbel (2002).

When it comes to estimation, we follow a semiparametric route. That is, we assume that the distribution of marginal effects follows a known parametric distribution governed by a finite parameter $\theta(x)$ which depends on the position $X_1 = X_2 = x$ at which we evaluate the conditional distribution. As such, our approach can be described as conditionally parametric. The advantage of such a procedure is as follows: Since our identification argument and the associated sample counterparts estimator is based on (conditional) characteristic functions, we avoid having to invert these estimators to obtain the (conditional) density. In the sample, this inversion step comes at the cost of having to pick an additional regularization parameter. Moreover, since one of the main objectives of our approach is to get an estimator for the quantiles of marginal effects as well, we avoid having to add another cumbersome inversion. Instead, the conditional parametric approach obtains all of these quantities: the conditional characteristic function, density, as well as the quantiles in one convenient step. Moreover, the characteristic function need not be observed for every value of the argument ($s$, say).

The core principle employed in our estimator is a minimum contrast step. We first form the sample counterpart to the identified nonparametric characteristic function for every value of $X_1 = X_2 = x$, and then pick the the parameter $\theta(x)$ that minimizes the contrast (distance) between the approximating parametric specification and this object. For this estimator, we establish the (optimal) minimax rate, and establish that our estimator achieves this rate. The rate is governed by the dimensionality of $X$ and the fact that we work with the set $X_1 = X_2 = x$. If there is no $U_{k,t}$ and $X$ is scalar, the rate is equivalent to a two dimensional nonparametric regression. Having, in addition, a $U_{k,t}$ that follows an ordinary smooth distribution slows the convergence rate down by the expected factor, $\alpha$, due to the added deconvolution step in removing the influence of $U_{k,t}$.

Importantly, this paper contains an application to consumer demand for junk food. Because of the relationship to obesity and other adverse health effects, this is a question of obvious importance for the society (see also the short literature review in the applied section). A key concern is that "poor" households - which we define to be households with low total expenditure for goods that Nielsen scanner data tracks - spend marginally more on junk food than wealthy, high income households. This means that a model that forces all households to have the same "income" and price elasticities, i.e., a linear random coefficients model, is not able to capture this important feature. Similarly, we want to control for unobserved factors that are correlated with poverty, e.g., education levels, in particular regarding nutrition, and hence it is imperative to allow for the unobservables to be correlated. Therefore, we feel that our approach, which allows for nonlinearities, high dimensional heterogeneity, and complicated correlation patterns, is particularly well suited for this application.

When applying our approach to the Nielsen Homescan data, we indeed find evidence of the aforementioned nonlinearities. Indeed, for every dollar spent on Nielsen products, poor households seem to consume twice as much junk food on average compared to wealthy households, even implicitly controlling for persistent correlated effects like education. Moreover, there also seems to be more heterogeneity within poor households (compared to wealthy ones), perhaps a function of the larger degree of addiction to an unhealthy lifestyle of at least parts of this subpopulation. It is interesting to muse about the reason for the significant correlation between expenditure levels and marginal effects, even after controlling for fixed factors. We also find very reasonable price elasticities that increase in the own price. Since we use a bundle of goods and Stone-Lewbel prices, we feel that this reflects heterogeneity in the composition of junk food. The more high level it is, the higher the price and the more elastic demand. More details can be found in the section on the application below.

**Related Literature:** Analyzing nonlinear panel data models has a long tradition, dating back to the conditional ML approach by Rasch (1961); see also Andersen (1970), Chamberlain (1982) and Chamberlain (1984) for models with non-additive individual heterogeneity. Nonlinear parametric panel data models have frequently been analyzed. For an overview of work related to discrete choice models, see Arellano (2003). Closely related to our work is that of Graham and Powell (2012), and Arellano and Bonhomme (2012), who consider estimation of moments and the distribution of random coefficients in a linear correlated coefficient panel data model. Compared to this line of work, we allow for the structural model to be arbitrarily nonlinear. Chamberlain (2010) discusses the identification of the dynamic panel data binary choice model, and why the logistic distribution assumption is required for identification of $\beta_o$, unless one is willing to assume unbounded support for one of the regressors, as is the case in Manski (1987). For other nonlinear fixed effects models, see also Hausman et al. (1984) for panel count data and Honoré (1992) for panel censored regression. Like all of this work, our approach assumes a fixed number of time periods. Indeed, it is one of the appealing features of our approach that we only require $T = 2$.

All of the work just described is concerned with a specific semiparametric model, e.g., the dynamic binary choice model. Approaches that are closer in spirit to our work are those of Chernozhukov et al. (2009), who consider discrete variation, whereas we consider derivatives, and Graham and Powell (2012), who focus on a linear heterogeneous population (i.e., the structure is linear in the coefficients, with coefficients that vary across the population) and not on a fully nonseparable structure. Other than the differences mentioned above, Graham and Powell (2012) also require (at least) as many time periods as regressors plus one, while we require only two time periods, even with a large number of regressors. Less closely related is the work on the correlated random coefficients models in panel data, see in particular Wooldridge (2005) and Murtazashvili and Wooldridge (2008). This line of work studies the linear random coefficients model as well, but imposes restriction on the correlation between time invariant individual specific effects and covariates of interest. In contrast, our approach allows for unobserved heterogeneity to enter nonlinearly and does not limit its correlation with the covariates of interest.

Finally, related is also the literature on nonseparable models using panel data, in particular Altonji and Matzkin (2005), Evdokimov (2010), Hoderlein and White (2012) and Chernozhukov et al. (2015). Unlike our paper, Altonji and Matzkin (2005) impose constraints on the correlation between $A_k$ and the $X_{k,t}$ process, but are more general in the structural function $\phi$ in that they allow interaction between the transitory error $U_{k,t}$ and the other variables, and focus on averages. Evdokimov (2010) also imposes additivity of the error $U_{k,t}$, but assumes that $A_k$ is a scalar and independent of $X_{k,t}$. Hoderlein and White (2012) and Chernozhukov et al. (2015) again admit a more general structural function $\phi$ (as in Altonji and Matzkin (2005)), but are only able to identify averages of the marginal effects, even though Chernozhukov et al. (2015) use distributional information. Instead, in this paper we use a deconvolution step to purge the model from the influence of $U_{k,t}$. This also allows to impose different, and arguably weaker, assumptions on the $U_{k,t}$-process. In particular, we do not require the stationarity assumption in their papers (Manski, 1987).

**Outline of the Paper:** Section 2 introduces the model and the precise assumptions we require. In Section 3, we present the general non-identification result for arbitrary values $x_2 \neq x_1$, which motivates our focus on the set of stayers. Section 4 then presents the main constructive nonparametric identification result and discusses extensions. Section 5 establishes the asymptotic lower bound for any estimator under this scenario. In Section 6, we introduce our conditional parametric estimator and the modeling assumptions, establish an upper bound under these conditions, and show that our estimator achieves the minimax rate. Section 7 analyzes the finite-sample performance of our estimators using several example of nonlinear heterogeneous DGPs. Section 8 discusses the application to consumer demand for junk food. The final section contains a summary and concluding remarks.

## 2.2 The Model: Basic Structure and Main Assumptions

We consider the panel data model

$$Y_{k,t} \;=\; \Phi(X_{k,t}, A_k) + U_{k,t}\,, \qquad k = 1, \ldots, n;\ , t = 1, \ldots, T\,, \tag{2.2.1}$$

where all $X_{k,t}$ and $Y_{k,t}$ are observed. Therein, the random vectors $(X_{k,t}, A_k, U_{k,t})_{t=1,\ldots,T}$ are i.i.d. (i.e. independent copies) for all $k = 1, \ldots, n$. Therefore, when addressing identification issues, we omit the index $k$ in the notation of all random variables. We impose the following assumptions:

(A.2.1) The random vectors $U := (U_1, \ldots, U_T)$ and $(A, X_1, \ldots, X_T)$ are independent.

This assumption is similar in spirit to the strict exogeneity assumptions commonly invoked in the panel data literature. It could be weakened, as is obvious from the proof. In particular, for $T = 2$ and using the notation $\Delta S = S_1 - S_2$ for any random variable $S$, we only need that $\Delta U$ independent of

$\Delta X, A|X_1$. However, since we use this stronger version in the construction of the estimator, we impose it henceforth.

(A.2.2) The random vector $X := (X_1, \ldots, X_T)$ has a $T$-dimensional Lebesgue density.

Our goal is to identify the conditional distribution $\mathcal{L}(Z_j \mid X)$ of the random variable

$$Z_j := \frac{\partial \Phi}{\partial x}(x, A) \,|_{x=X_j} \,,$$

given $X$. From a famous result in probability theory (e.g. p. 439, Theorem 33.3, Billingsley (1995)), we learn that there exists a function $\zeta_j$ from the domain $\mathbb{R}^T$ to the set of all probability measures on the Borel $\sigma$-field $\mathfrak{B}(\mathbb{R})$ of $\mathbb{R}$ such that

$$\big\{\zeta_j(X)\big\}(B) = P[Z_j \in B \mid X] \,, \text{ a.s.} \,,$$

for all elements $B$ of the Borel $\sigma$-field $\mathfrak{B}(\mathbb{R})$. This equation, however, does not determine the value of the mapping $\zeta_j$ at any fixed $x \in \mathbb{R}^T$. In particular, the value of $\zeta_j$ at one singular $x \in \mathbb{R}^T$ can be changed without switching to an observationally non-equivalent model due to condition (A2). As a consequence, identification and estimation of $\zeta_j(x)$, for any specific value $x \in \mathbb{R}^T$, is impossible unless continuity conditions are assumed such as

(A.2.3) There exists a function $\zeta_j$ on the domain $\mathbb{R}^T$ to the set of all probability measures on $\mathfrak{B}(\mathbb{R})$ which is continuous with respect to the Fourier distance on its codomain; and satisfies

$$\big\{\zeta_j(X)\big\}(B) = P[Z_j \in B \mid X] \,, \text{ a.s.} \,,$$

for all $B \in \mathfrak{B}(\mathbb{R})$.[1]

[1] Here, the Fourier distance between two probability measures $P$ and $Q$ on $\mathfrak{B}(\mathbb{R})$ is defined

Condition (A.2.3) resembles the usual constraints in the setting of standard nonparametric regression where the regression function is required to be continuous under continuously distributed covariates in order to attain pointwise consistency at a fixed site. The following lemma shows that $\zeta_j(x)$ is uniquely determined for each $x$ in the support of $X$.

**Lemma 2.2.1.** *Assume two functions $\zeta_j$ and $\tilde{\zeta}_j$ which satisfy the continuity assumptions imposed on $\zeta_j$ in (A.2.3); and*

$$\big\{\zeta_j(X)\big\}(B) \;=\; P[Z_j \in B \mid X] \;=\; \big\{\tilde{\zeta}_j(X)\big\}(B) \;\; a.s., \forall B \in \mathfrak{B}(\mathbb{R})\,.$$

*Then the restrictions of $\zeta_j$ and $\tilde{\zeta}_j$ to the support $\mathcal{S}_X$ of $X$ coincide.*

## 2.3 Non-Identification

Now we focus on the question for which elements $x$ of $\mathcal{S}_X$ the probability measure $\zeta_j(x)$ can be identified from the observed data $(X_t, Y_t)$, $t = 1, \ldots, T$, under the Assumptions (A.2.1)–(A.2.3). Using the notation $p(x) := (1, x^1, \ldots, x^T)^\dagger$ and $q(x) := (0, 1, 2x, \ldots, Tx^{T-1})^\dagger$, we provide the following useful tool.

**Lemma 2.3.1.** *The vectors $p(x_1), \ldots, p(x_T), q(x_j)$, for any $j \in \{1, \ldots, T\}$, are linearly independent if and only if all $x_1, \ldots, x_T$ differ from each other.*

By $H(x)$, $x = (x_1, \ldots, x_T)$, we denote the linear hull of $p(x_1), \ldots, p(x_T)$. The squared distance between $H(x)$ and $q(x_j)$ is called $\tau_j(x)$.

by

$$\mathcal{F}(P,Q) \;:=\; \sup_{s\in\mathbb{R}} \big|P^{ft}(s) - Q^{ft}(s)\big|\,, \tag{2.2.2}$$

where $P^{ft}(s) := \int \exp(isx)dP(x)$ denotes the Fourier transform of $P$. Note that the total variation distance $\mathrm{TV}(P,Q)$ between $P$ and $Q$, i.e.

$$\mathrm{TV}(P,Q) \;:=\; \sup_{B\in\mathfrak{B}(\mathbb{R})} |P(B) - Q(B)|\,,$$

dominates the Fourier distance $\mathcal{F}(P,Q)$. The set of all probability measures on $\mathfrak{B}(\mathbb{R})$, equipped with the Fourier distance $\mathcal{F}$, forms a complete metric space thanks to the completeness of the space $C_0(\mathbb{R})$ and Lévy's continuity theorem (e.g. Williams, 1991, section 18.1).

**Lemma 2.3.2.** *The function $\tau_j$ is continuous and takes on only strictly positive values on the set $\mathcal{T}_X := \bigcap_{k \neq l} \{x \in \mathbb{R}^T : x_k \neq x_l\}$.*

In order to prove a non-identification result, we may, in addition, assume that the function $\Phi$ and the distribution of the random vector $U$, as well as the distribution of the covariates $X$, are known. Concretely, we impose that

$$\Phi(x, A) \;=\; \sum_{t=0}^{T} A_t x^t \,. \tag{2.3.1}$$

Let $q_j^*(x)$ denote the orthogonal projection of $q(x_j)$ onto the orthogonal complement of $H(x)$ with respect to $\mathbb{R}^{T+1}$ as this notation has already been used in the proof of Lemma 2.3.2. This lemma also yields $q_j^*(x) \neq 0$ for all $x \in \mathcal{T}_X$ since $|q_j^*|^2 = \tau_j$. Then we are ready to define the random variables

$$A^{[b]} \;:=\; A^{[0]} \;+\; \sqrt{b}\,\delta\, q_j^*(X)\,, \qquad b \geq 0\,, \tag{2.3.2}$$

where the random variable $\delta$ is standard normal; $A^{[0]}$ is an arbitrary $(T+1)$-dimensional random vector; and $(X, A^{[0]})$ and $\delta$ are independent. Then

$$\mathcal{L}^{[b]}(A \mid X) = \mathcal{L}(A^{[b]} \mid X)\,, \qquad b \geq 0\,,$$

denote competing candidates for the conditional distribution of $A$ given $X$.

The conditional characteristic function of $V := \big(\Phi(X_1, A), \ldots, \Phi(X_T, A)\big)$ given $X$ equals

$$\begin{aligned}\psi_{V|X}(t) \;&=\; E\Big\{\exp\Big(i\sum_{k=1}^{T} t_k \Phi(X_k, A)\Big) \mid X\Big\} \;=\; E\Big\{\exp\Big(i\sum_{l=0}^{T} A_l \sum_{k=1}^{T} t_k X_k^l\Big) \mid X\Big\} \\ &=\; \psi_{A|X}\Big(\sum_{k=1}^{T} t_k X_k^0, \ldots, \sum_{k=1}^{T} t_k X_k^T\Big)\,,\end{aligned}$$

for all $t \in \mathbb{R}^T$ whenever (2.3.1) holds true. Hence, for the candidates $\mathcal{L}^{[b]}(A \mid X)$,

$b \geq 0$, it holds that

$$\begin{aligned}\psi^{[b]}_{V|X}(t) &= \exp\Big(-\frac{1}{2}b\Big|\sum_{k=1}^{T} t_k p(X_k)^\dagger q_j^*(X)\Big|^2\Big) \cdot \psi_{A^{[0]}|X}\Big(\sum_{k=1}^{T} t_k p(X_k)\Big) \\ &= \psi_{A^{[0]}|X}\Big(\sum_{k=1}^{T} t_k p(X_k)\Big),\end{aligned}$$

for all $t \in \mathbb{R}^T$ and $b \geq 0$ so that the conditional distributions $\mathcal{L}^{[b]}(V \mid X)$ coincide almost surely for all $b \geq 0$. Therefore, the distribution of the observed data $(X, Y)$ with $Y := (Y_1, \ldots, Y_T)$, are identical for all candidates $(b \geq 0)$ thanks to the independence of $U$ and $(A, X)$. Therefore one is unable to decide what is the value of $b$ based on the distribution of the observations.

Due to (2.3.1) and (2.3.2) we have $Z_j^{[b]} = (A^{[b]})^\dagger q(X_j)$ so that

$$\zeta_j^{[b]}(x) = \mathcal{L}\big({A^{[0]}}^\dagger q(x_j) \mid X = x\big) * \mathrm{N}(0, b\tau_j^2(x)), \tag{2.3.3}$$

where $*$ denotes convolution. Consider $\mathrm{N}(0,0)$ as the Dirac measure which is concentrated at 0. The corresponding Fourier transform equals

$$\big\{\zeta_j^{[b]}(x)\big\}^{ft}(s) = \psi_{A^{[0]}|X=x}\big(sq(x_j)\big) \cdot \exp\Big(-\frac{1}{2}bs^2\tau_j^2(x)\Big), \qquad s \in \mathbb{R}. \tag{2.3.4}$$

We impose the Assumption

(A.2.4) The random vector $A^{[0]}$ has a conditional Lebesgue density $f_{A^{[0]}|X=x}$ given $X = x$ for all $x \in \mathbb{R}^T$; moreover, we have that

$$\lim_{y \to x} \mathcal{F}\big(\mathcal{L}(A^{[0]} \mid X = x), \mathcal{L}(A^{[0]} \mid X = y)\big) = 0, \qquad \forall x \in \mathbb{R}^T.$$

In Assumption (A.2.4), we have extended the definition of the Fourier distance in (2.2.2) to probability measures on $\mathfrak{B}(\mathbb{R}^{T+1})$ in a natural way by the supremum norm distance of the Fourier transforms of both measures. Note that Assumption (A.2.4) is satisfied in particular if $A^{[0]}$ has a Lebesgue density and $A^{[0]}$ and $X$ are independent, which is related to the scenario considered in Evdokimov (2010). The following lemma verifies Assumption (A.2.3) in our setting.

**Lemma 2.3.3.** *The functions $\zeta_j^{[b]}$ in (2.3.3) are continuous for any $b \geq 0$ with respect to the Fourier distance on the codomain under the Assumption (A.2.4).*

Furthermore Lemma 2.3.2 and the equation (2.3.3) yield that, for all $b \neq b' > 0$, the probability measures $\zeta_j^{[b]}(x)$ and $\zeta_j^{[b']}(x)$ are different from each other for all $x \in \mathcal{S}_X \cap \mathcal{T}_X$ where we use the following result.

**Lemma 2.3.4.** *Let $Q$ be an arbitrary probability measure on $\mathfrak{B}(\mathbb{R})$. Then the equality $Q * N(0, \alpha) = Q * N(0, \alpha')$ implies $\alpha = \alpha'$ for all $\alpha, \alpha' \in [0, \infty)$.*

Thus, we have established the following theorem about non-identification of $\zeta_j(x)$, for all $x \in \mathcal{S}_X \cap \mathcal{T}_X$, i.e., values of $x$ for which $x_1 \neq x_2$, in the model (2.2.1).

**Theorem 2.1.** *In the model (2.2.1), fix some $j = 1, \ldots, T$; select the function $\Phi$ as in (2.3.1); and grant the Assumptions (A.2.1) and (A.2.2). Set the random variable $A$ equal to $A^{[b]}$ in (2.3.2) where the choice of $A^{[0]}$ is only restricted by Assumption (A.2.4). Then the corresponding distributions of the observations $(X, Y)$ coincide for all $b \geq 0$ while Assumption (A.2.3) is satisfied for all $b \geq 0$; and $\zeta_j^{[b]}(x) \neq \zeta_j^{[b']}(x)$ holds true for all $b \neq b'$ and $x \in \mathcal{S}_X \cap \mathcal{T}_X$.*

## 2.4 Identification

Now assume that $T = 2$ and $j = 1$. According to Theorem 2.1, the function $\zeta(x)$ cannot be identified from the data distribution unless we restrict to $x \in \mathcal{S}_X \backslash \mathcal{T}_X$, which equals $\{(x_1, x_2) \in \mathcal{S}_X : x_1 = x_2\}$. Moreover we impose

(A.2.5) There exists some $\rho > 0$ such that the density $f_X$ of $X = (X_1, X_2)$ is continuous and strictly positive on the strip

$$\mathcal{S}_X^{(\rho)} := \{(x_1, x_2) \in \mathbb{R}^2 : |x_1 - x_2| \le \rho\}\,.$$

Under Assumption (A.2.5) it holds that $\mathcal{S}_X \backslash \mathcal{T}_X$ is a subset of $\mathcal{S}_X^{(\rho)}$. The smoothness condition (A.2.4) is quantified via the Assumption

(A.2.6) The function $\Phi$ is twice continuously differentiable and we have

$$E\Big( \sup_{\xi \in [X_1, X_2] \cup [X_2, X_1]} \Big| \frac{\partial^j \Phi}{\partial x^j}(\xi, A) \Big| \, \Big| \, X_1, X_2 \Big) \le c_\Phi \qquad \text{a.s.,}$$

for $j = 1, 2$ and some constant $c_\Phi$. Moreover $\zeta_1$ satisfies the Lipschitz condition

$$\mathcal{F}\big(\zeta_1(x), \zeta_1(y)\big) \le c_\zeta |x - y|\,, \qquad \forall x, y \in \mathcal{S}_X^{(\rho)}\,,$$

for some constant $c_\zeta \in (0, \infty)$.

We introduce the notation

$$\begin{aligned} \Delta Y &:= Y_1 - Y_2 = \Delta\Phi + \Delta U\,, \\ \Delta\Phi &:= \Phi(X_1, A) - \Phi(X_2, A)\,, \\ \Delta U &:= U_1 - U_2\,, \\ \Delta X &:= X_1 - X_2\,. \end{aligned} \tag{2.4.1}$$

The Assumption saying that

(A.2.1') the random variables $\Delta\Phi$ and $\Delta U$ are conditionally independent given $X$; and $\Delta U$ and $X$ are independent,

is weaker than the Assumption (A.2.1) and suffices to show that the corresponding conditional characteristic functions satisfy

$$\psi_{\Delta Y|X} \;=\; \psi_{\Delta\Phi|X} \cdot \psi_{\Delta U}\,, \qquad \text{a.s.}\,. \tag{2.4.2}$$

In fact, as mentioned before and as is obvious from what follows, this assumption could even be weakened further, but since we implement our estimator with this stronger assumption we desist from doing so. For some $h_0 \in (0, \rho)$ let us consider the term

$$\begin{aligned} T_U(h_0, s) &:= E\exp(is\Delta Y) \cdot 1_{\mathcal{S}_X^{(h_0)}}(X)/P\big[X \in \mathcal{S}_X^{(h_0)}\big] \\ &\;=\; E\psi_{\Delta Y|X}(s) \cdot 1_{\mathcal{S}_X^{(h_0)}}(X)/P\big[X \in \mathcal{S}_X^{(h_0)}\big] \\ &\;=\; \psi_{\Delta U}(s) \cdot E\psi_{\Delta\Phi|X}(s) \cdot 1_{\mathcal{S}_X^{(h_0)}}(X)/P\big[X \in \mathcal{S}_X^{(h_0)}\big]\,, \end{aligned}$$

for any $s \in \mathbb{R}$, which is directly accessible from the distribution of the observation $(X, Y)$. Therein note that $P\big[X \in \mathcal{S}_X^{(h_0)}\big] > 0$ is guaranteed for any $h_0 \in (0, \rho)$ by Assumption (A.2.5); and that we have used (2.4.2). By Assumption (A.2.6) it holds that

$$\begin{aligned} \big|E\psi_{\Delta\Phi|X}(s) \cdot 1_{\mathcal{S}_X^{(h_0)}}(X) \;-\; P\big[X \in \mathcal{S}_X^{(h_0)}\big]\big| \;&\leq\; c_\Phi\,|s|\,E|\Delta X| \cdot 1_{\mathcal{S}_X^{(h_0)}}(X) \\ &\leq\; c_\Phi\,|s|\,h_0\,P\big[X \in \mathcal{S}_X^{(h_0)}\big]\,, \end{aligned} \tag{2.4.3}$$

so that

$$\big|T_U(h_0, s) - \psi_{\Delta U}(s)\big| \;\leq\; c_\Phi\,|s|\,|\psi_{\Delta U}(s)|\,h_0\,,$$

and, thus, $\lim_{h_0\downarrow 0} T_U(h_0, s) = \psi_{\Delta U}(s)$ for any $s \in \mathbb{R}$. Therefore $\psi_{\Delta U}$ and, hence, the distribution of $\Delta U$ are identified from the distribution of $(X, Y)$. This motivates the following estimator of $\psi_{\Delta U}(s)$,

$$\hat{\psi}_{\Delta U}^{(h_0)}(s) \;:=\; \sum_{k=1}^{n} \exp\big(is\Delta Y_k\big) \cdot 1_{\mathcal{S}_X^{(h_0)}}(X_{k,1}, X_{k,2}) \,\Big/\, \sum_{k=1}^{n} 1_{\mathcal{S}_X^{(h_0)}}(X_{k,1}, X_{k,2})\,, \tag{2.4.4}$$

based on the moment method, for some $h_0 \in (0,\rho)$ still to be selected. By convention put $\hat{\psi}_{\Delta U}^{(h_0)}(s)$ equal to 0 if the denominator in (2.4.4) vanishes.

Writing $\mathcal{S}_X^{(h_1,h_2)} := \mathcal{S}_X^{(h_2)} \backslash \mathcal{S}_X^{(h_1)}$ for some $\rho > h_2 > h_1 > 0$, we consider the term

$$\begin{aligned} T_Z := T_Z(x,h_1,h_2,h_3,s) := & \; E\psi_{\Delta U}(-s/\Delta X)\exp(is\Delta Y/\Delta X)1_{\mathcal{S}_X^{(h_1,h_2)}}(X)1_{[0,h_3]}(|X_1 - x|) \\ & \; /E\big|\psi_{\Delta U}(s/\Delta X)\big|^2 1_{\mathcal{S}_X^{(h_1,h_2)}}(X)1_{[0,h_3]}(|X_1 - x|) \\ = & \; E\big|\psi_{\Delta U}(s/\Delta X)\big|^2 \psi_{\Delta\Phi|X}(s/\Delta X)1_{\mathcal{S}_X^{(h_1,h_2)}}(X)1_{[0,h_3]}(|X_1 - x|) \\ & \; /E\big|\psi_{\Delta U}(s/\Delta X)\big|^2 1_{\mathcal{S}_X^{(h_1,h_2)}}(X)1_{[0,h_3]}(|X_1 - x_1|)\,, \end{aligned}$$

for some $h_3 > 0$ and any fixed $x = (x_1, x_2)$ with $x_1 = x_2$, which is directly accessible from the distribution of $(X,Y)$ as $\psi_{\Delta U}$ has already been identified. Again we have used (2.4.2). Combining Assumption (A.2.5) with the Assumption

(A.2.7) The characteristic function $\psi_{\Delta U}$ does not vanish,

we may ensure that the denominator of the term $T_Z$ does not vanish. Assumption (A.2.6) and Taylor approximation yield that

$$\Delta\Phi \;=\; Z_1 \cdot \Delta X \;+\; \mathcal{R}\,, \tag{2.4.5}$$

where the random remainder term $\mathcal{R}$ satisfies

$$\big|\mathcal{R}\big| \;\le\; \frac{1}{2}c_\Phi(\Delta X)^2 \qquad \text{a.s.}\,.$$

It follows from there that, on the event $\big\{X \in \mathcal{S}_X^{(h_1,h_2)}\big\} \cap \{|X_1 - x_1| \le h_3\}$, we have that

$$\big|\psi_{\Delta\Phi|X}(s/\Delta X) - \{\zeta_1(x)\}^{ft}(s)\big| \;\le\; \big(c_\Phi|s|/2\big)\,h_2 \;+\; c_\zeta\,(2h_3 + h_2)\,,$$

using Assumption (A.2.6) so that

$$\lim_{h_2 \downarrow 0} T_Z(x, h_1, h_2, h_3, s) \;=\; \{\zeta_1(x)\}^{ft}(s)\,,$$

for all $s \in \mathbb{R}$ where we arrange that $h_1 = h_2/2$ and $h_3 = h_2$ to calculate the limit. Note that $x \in \bigcap_{h_2>0} \mathcal{S}_X^{(h_2/2,h_2)}$. Therefore $\zeta_1(x)$ is identified. Moreover the quantity $T_Z$ along with its asymptotic behavior motivates an estimator of $\{\zeta_1(x)\}^{ft}(s)$, namely

$$\begin{aligned}
&\hat{\psi}_{Z_1}^{(h_0,h_1,h_2,h_3)}(x;s) \\
&:= \sum_{k=1}^{n} \exp\big(is\Delta Y_k/\Delta X_k\big)\hat{\psi}_{\Delta U}^{(h_0)}(-s/\Delta X_k)\cdot 1_{\mathcal{S}_X^{(h_1,h_2)}}(X_{k,1},X_{k,2})\cdot K(|X_{k,1}-x_1|/h_3) \\
&\qquad / \Big\{\rho_n + \sum_{k=1}^{n} \big|\hat{\psi}_{\Delta U}^{(h_0)}(s/\Delta X_k)\big|^2 \cdot 1_{\mathcal{S}_X^{(h_1,h_2)}}(X_{k,1},X_{k,2})\cdot K(|X_{k,1}-x_1|/h_3)\Big\}\,,
\end{aligned} \tag{2.4.6}$$

for some $0 < h_0 < h_1 < h_2 < \rho$, $h_3 > 0$, some kernel function $K$ and some ridge parameter $\rho_n > 0$ in order to prevent the denominator from getting too close to zero. This approach to heteroskedastic deconvolution is inspired by Delaigle and Meister (2007); Delaigle et al. (2008).

Before studying the estimator (2.4.6) let us summarize the identification result in the following theorem.

**Theorem 2.2.** *Under the Assumptions (A.2.1'), (A.2.2), (A.2.3) and (A.2.5)–(A.2.7), $\zeta_1(x)$ is identified in the model (2.2.1) for any $x = (x_1, x_2) \in \mathbb{R}^2$ with $x_1 = x_2$ from the distribution of the observations $(X, Y)$.*

**Remark.** The model (2.2.1) may be generalized to the setting of multiple regressors, i.e. one observes the i.i.d. data $(X_{k,t}, X'_{k,t}, Y_{k,t})$, $k = 1, \ldots, n$, $t = 1, 2$, where

$$Y_{k,t} \;=\; \Phi(X_{k,t}, X'_{k,t}, A_k) + U_{k,t}\,.$$

Then we modify the definition

$$Z_j := \frac{\partial \Phi}{\partial x}(x, x', A) \,|_{x=X_j, x'=X'_j} \,,$$

and that of $\zeta_j$ accordingly. Let us assume that $(X_{1,1}, X'_{1,1}, X_{1,2}, X'_{1,2})$ has a four dimensional Lebesgue density, which is continuous and strictly positive. Also impose additional Lipschitz conditions on $\zeta_1$ and its partial derivatives with respect to the bivariate component $(x, x')$ in Assumption (A.2.6). Then Theorem 2.2 can be extended to identify $\zeta_1(x, x')$ at any $(x, x') = (x_1, x_2, x'_1, x'_2)$ with $x_1 = x_2$ and $x'_1 = x'_2$. For any unitary matrix $U$ define

$$\tilde{\Phi}(x, y, a) \;:=\; \Phi(U^T(x, y)^T, a)\,.$$

Then use the above arguments to identify the conditional distribution of

$$\frac{\partial \tilde{\Phi}}{\partial x}(W_1, W'_1, A) \;=\; U_{1,1} \cdot \frac{\partial \Phi}{\partial x}(X_1, X'_1, A) + U_{1,2} \cdot \frac{\partial \Phi}{\partial x'}(X_1, X'_1, A)\,,$$

given $W_1 = W_2$ and $W'_1 = W'_2$ based on the data $Z_j$ and $(W_j, W'_j)^T = U(X_j, X'_j)^T$ for $j = 1, 2$. That opens the perspective to identify any directional derivative of $\Phi$ at $x_1 = x_2 = x$ and $x'_1 = x'_2 = x'$ and, hence, the gradient of $\Phi$ under appropriate smoothness conditions on $\Phi$ and $\zeta_1$.

**Remark.** If there are more time periods, it is also possible to allow for a time trend. Specifically, we allow for a linear time trend which modifies the structural function $\phi$ by adding the same the structural function in each time period. More formally, the model takes the form

$$Y_{k,t} = \Phi_0\left(X_{k,t}, A_k\right) + \Phi_1\left(X_{k,t}, A_k\right) t + U_{kt}, \quad t = 1, \ldots, T, \quad k = 1, \ldots, n, \qquad (2.4.7)$$

where $\Phi_0$ and $\Phi_1$ satisfy analogous conditions to before. To identify this model, we require $T = 4$. Since

$$\begin{aligned} Y_{1,2} - Y_{1,1} &= U_{1,2} - U_{1,1} + \Phi_1(X_{1,1}, A_1)\,, \\ Y_{1,4} - Y_{1,3} &= U_{1,4} - U_{1,3} + \Phi_1(X_{1,3}, A_1)\,, \end{aligned}$$

holds on the event $\{X_{1,1} = X_{1,2}, X_{1,3} = X_{1,4}\}$ we are able to identify the conditional distribution of $\partial_x \Phi_1(x, A) \mid_{x=X_{1,1}}$ given $X_{1,1} = x$ at $x = \lambda \cdot (1,1,1,1)$, $\lambda \in \mathbb{R}$, by the arguments from section 2.4 under the given assumptions. Moreover

$$\begin{aligned} 2Y_{1,1} - Y_{1,2} &= 2U_{1,1} - U_{1,2} + \Phi_0(X_{1,1}, A_1)\,, \\ 4Y_{1,3} - 3Y_{1,4} &= 4U_{1,3} - 3U_{1,4} + \Phi_0(X_{1,3}, A_1)\,, \end{aligned}$$

holds on $\{X_{1,1} = X_{1,2}, X_{1,3} = X_{1,4}\}$ again so that the conditional distribution of $\partial_x \Phi_0(x, A) \mid_{x=X_{1,1}}$ given $X_{1,1} = x$ at $x = \lambda \cdot (1,1,1,1)$, $\lambda \in \mathbb{R}$, is identified as well. Note that continuity conditions analogous to Assumption (A.2.6) have to be imposed on both $\Phi_0$ and $\Phi_1$.

**Remark.** Our framework may be extended to allow for additional covariates, denoted in the following by $S_t$. The main motivation to do so stems typically from the objective to simply control for these variables; their influence is typically of lesser interest. Due to the curse of dimensionality, it is impractical to let them enter in an unrestricted fashion. Hence we propose a partially linear structure, i.e.,

$$Y_{k,t} = \Phi\left(X_{k,t}, A_k\right) + \gamma' S_{k,t} + U_{kt}, \quad t = 1, \ldots, T, \quad k = 1, \ldots, n, \tag{2.4.8}$$

where $\gamma \in R^{\dim(S_t)}$ is a fixed parameter. Constructive identification of $\gamma$ is straightforwardly established by noting that, conditional on $X_{k,1} = X_{k,2} = x$, this equation is

$$Y_{k,t} = \tilde{A}_k + \gamma' S_{k,t} + U_{kt}, \quad t = 1, \dots, T, \quad k = 1, \dots, n, \tag{2.4.9}$$

where $\tilde{A}_k = \Phi\left(x, A_k\right)$ is a classical, time invariant, additive "fixed effect". This implies that, for every value of $x$, we obtain a classical linear fixed effect model. Since the coefficient $\gamma$ is invariant over $x$, we can then average out over $x$. A sample counterpart estimator to this identification argument would produce an estimator that converges at the dim$(X)$ nonparametric regression rate (because we have to impose that $X_{k,1} = X_{k,2}$).

Finally, after forming $Y_{k,t} - \gamma' S_{k,t}$, the further analysis can proceed exactly as outlined above.

## 2.5 Asymptotic Lower Bound

In this section, we investigate the limits for the asymptotic performance of an arbitrary estimator under the conditions of Theorem 2.2. For that purpose we consider the polynomial approach (2.3.1) with $T = 2$ and the random vector $A$ equals

$$A \;=\; \begin{pmatrix} X_1 X_2 - (X_1 + X_2) B/2 \\ B - X_1 - X_2 \\ 1 \end{pmatrix}, \tag{2.5.1}$$

where the random vector $B$ remains to be specified. Under given $X = (X_1, X_2)$, observing

$$\begin{aligned} Y_1 + Y_2 \;&=\; U_1 + U_2\,, \\ \Delta Y/\Delta X \;&=\; B + \Delta U/\Delta X\,, \end{aligned} \tag{2.5.2}$$

is equivalent with the observation of the data $(Y_1, Y_2)$, i.e. the random variable $(Y_1, Y_2)$ can be uniquely reconstructed from (2.5.2) and vice versa. Then $\zeta_1(x)$, at any $x = (x_1, x_2)$ with $x_1 = x_2$, equals the conditional distribution of $B$ given $X = x$. With respect to the random vector $U$ we impose Assumption (A.2.1) and

(A.2.8) $U = (U_1, U_2)$ has the bivariate Lebesgue density

$$(s,t) \mapsto 2 f_{\Delta U}(s-t) f_{\Delta U}(s+t)\,,$$

where the Fourier transform of the univariate density $f_{\Delta U}$ satisfies

$$0 < c_{U,1} \leq (1+|t|^{\alpha}) \cdot \big|\psi_{\Delta U}(t)\big| \leq c_{U,2} < \infty\,, \qquad \forall t \in \mathbb{R}\,,$$

for some constants $\alpha > 0$ and $c_{U,1} < c_{U,2}$. Moreover $\psi_{\Delta U}$ is twice continuously differentiable and its derivatives satisfy

$$\sup_t\,(1+|t|^{\alpha+\ell}) \cdot \big|\psi_{\Delta U}^{(\ell)}(t)\big| \leq c_{U,3}\,,$$

for another constant $c_{U,3} > 0$ and $\ell = 1, 2$.

Under the Assumption (A.2.8), $f_{\Delta U}$ is an ordinary smooth density in the terminology of Fan (1991). Moreover (A.2.8) yields that $U_1 + U_2$ and $\Delta U$ are independent and that $\Delta U$ has the density $f_{\Delta U}$. Considering (2.5.2), it follows that

$$(X_{j,t}, \Delta Y_j/\Delta X_j), \qquad j = 1, \ldots, n,\ t = 1, 2\,, \tag{2.5.3}$$

forms a sufficient statistic for $\zeta_1(x)$ in the model in which the data $(X_{j,t}, Y_{j,t})$, $j = 1, \ldots, n$, $t = 1, 2$, are observed. Therefore we may focus on that experiment in which only the i.i.d. sample (2.5.3) is available.

Let us now determine the conditional distribution of $B$ given $X$. Define

$$f_0(x) := c \cdot \{1 - \cos(x)\}^2/x^4\,, \qquad x \in \mathbb{R}\,,$$

with some constant $c > 0$ such that $f_0$ integrates to one. We introduce

$$f_{B|X}^{[\theta]}(t) \;:=\; \frac{3}{4}\cdot(1+|t|)^{-4} + \frac{1}{2}f_0(t)\cdot\{1+\theta\cdot K\big(|X-x|/\theta\big)\cdot\cos(4t)\big\}\,, \qquad \forall t \in \mathbb{R}\,, \tag{2.5.4}$$

for any $\theta \in [0,1]$, as the competing conditional densities of $B$ given $X$. Therein $K$ denotes some continuously differentiable kernel function which is supported on $[-1,1]$, bounded by 1 and satisfies $K(0) = 1$. As $f_0^{ft}$ is supported on $[-2,2]$ the function $f_{B|X}^{[\theta]}$ is a probability density indeed. Moreover we put $f_{B|X}^{[0]}(t) := 3(1+|t|)^{-4}/4 \,+\, f_0(t)/2$.

With respect to the design distribution we modify Assumption (A.2.5) via

(A.2.5') There exists some $\rho > 0$ such that the density $f_X$ of $X = (X_1, X_2)$ is continuous and strictly positive on the ball around $x = (x_1, x_1)$ with the radius $\rho$. Moreover $f_X$ is compactly supported.

We provide the following lower bound on the convergence rates for the estimation of the parameter $\theta$ in the model (2.5.4).

**Theorem 2.3.** *We impose that $\Phi$ has the polynomial shape (2.3.1) with $T = 2$; that $A$ and $B$ obey (2.5.1) and (2.5.4), respectively; and that the Assumptions (A.2.1), (A.2.2), (A.2.5') and (A.2.8) hold true. Then Assumption (A.2.6) is satisfied for appropriate finite constants $c_\Phi$ and $c_\zeta$. For an arbitrary sequence of estimators $(\hat\theta_n)_n$, where $\theta_n$ is based on the i.i.d. data $(X_{j,t}, Y_{j,t})$, $j = 1,\ldots,n$, $t = 1,2$, there exists a constant $d > 0$ such that*

$$\liminf_{n\to\infty} \sup_{\theta\in[0,1]} P_\theta^{(n)}\big[|\hat\theta_n - \theta|^2 \;>\; d^2\cdot n^{-1/(2+\alpha)}\big] \;>\; 0\,.$$

## 2.6 A Conditional Parametric Estimator

In this section, our goal is to construct a parametric estimator of $\zeta_1(x)$ which attains the convergence rates outlined in Theorem 2.3. The parametric nature of the estimation problem is represented by the following assumption

(A.2.9) For some fixed $x = (x_1, x_2) \in \mathbb{R}^2$ with $x_1 = x_2$, there exists a parametrization

$$\theta \in \Theta \subseteq \mathbb{R}^d,\ \theta \mapsto \zeta_1(\theta; x)\,,$$

of the admitted conditional measures $\zeta_1(x)$ for $d \geq 1$ such that

$$\inf_{\theta' \neq \theta \in \Theta} \mathcal{F}_R\big(\zeta_1(\theta'; x), \zeta_1(\theta; x)\big)/|\theta' - \theta| \;\geq\; c_p \;>\; 0\,,$$

holds true for some fixed $R \in (0, \infty)$.

Therein $\mathcal{F}_R$ denotes following distance between two probability measures $P$ and $Q$,

$$\mathcal{F}_R^2(P, Q) \;:=\; \int_{-R}^{R} \big|P^{ft}(t) - Q^{ft}(t)\big|^2 dt\,.$$

The specific parametrization in (2.5.4), which has been used to prove the lower bound in Theorem 2.3, satisfies Assumption (A.2.9) when putting

$$c_p^2 \;=\; \frac{\pi}{8} \int f_0^2(t) dt\,.$$

As the estimator $\hat{\theta}$ of $\theta$ we define that $\tilde{\theta}$ which minimizes the contrast functional

$$\gamma(x; \tilde{\theta}) \;:=\; \int_{-R}^{R} \big|\hat{\psi}_{Z_1}^{(h_0,h_1,h_2,h_3)}(x; s) - \{\zeta_1(\tilde{\theta}; x)\}^{ft}(s)\big|^2 ds\,,$$

among all $\tilde{\theta} \in \Theta$ where $\hat{\psi}_{Z_1}^{(h_0,h_1,h_2,h_3)}$ is as in (2.4.6) and $h_0$, $h_1$, $h_2$ and $h_3$ remain to be selected.

The following theorem provides an upper bound on the estimation error of our estimator $\hat{\theta}$ under appropriate selection of the smoothing parameters. For simplicity we restrict to the uniform kernel $K$.

**Theorem 2.4.** *We consider the model (2.2.1) for $T = 2$ under the Assumptions (A.2.1'), (A.2.2), (A.2.5'), (A.2.6), (A.2.8) and (A.2.9). The distribution of $(X_1, X_2)$ and the constants in the assumptions are imposed to be fixed while $\Phi$, $\theta$ and the distributions of $A$ and $(U_1, U_2)$ may move in $n$ and $d$. Then, our estimator $\hat{\theta}$ of $\theta$ satisfies*

$$\left|\hat{\theta} - \theta\right|^2 = \mathcal{O}_P\left(n^{-1/(2+\alpha)}\right),$$

*under the selection $K = 1_{[0,1]}$, $\rho_n \asymp 1$, $h_2 = 2h_1$, $h_3 \asymp h_1$, $h_0 \asymp h_1^2$, $h_1 \asymp n^{-1/(4+2\alpha)}$.*

Combining Theorem 2.3 and 2.4, it follows that our estimator $\hat{\theta}$ achieves the optimal minimax convergence rate. It is remarkable that, in spite of the parametric nature of the estimation problem, the usual square-root-asymptotics are not attainable by any estimator. In the error-free case (i.e. $\alpha = 0$), the convergence rate is $\mathcal{O}_P(n^{-1/4})$ with respect to the non-squared estimation error.

Critically we mention that the asymptotic order of $h_1$ in Theorem 2.4 depends on the parameter $\alpha$ from Assumption (A.2.8), which is usually unknown. Therefore we propose a data-driven choice of $h_1$ (and $h_0$, $h_2$, $h_3$ according to Theorem 2.4) by splitting the sample. Precisely the estimator $\hat{\theta}$ is only based on $\lfloor qn \rfloor$ of the complete sample for some constant $q \in (0, 1)$. All other observations are used to construct an empirical selector $\hat{h}_1$ of $h_1$ as follows: Define

$$\hat{\alpha} := -\left(\log\left|\hat{\psi}_{\Delta U}^{(h_4)}(s_n)\right|\right) / \log s_n,$$

with some deterministic positive parameters $h_4$ and $s_n > 1$ and the estimator of $\psi_{\Delta U}$ from (2.4.4); and, finally,

$$\hat{h}_1 := n^{-1/(4+2\hat{\alpha})}. \tag{2.6.1}$$

The following result suffices to show that the asymptotic upper bound from Theorem 2.4 is maintained when using the split-of-the-sample estimator with the plug-in selector $\hat{h}_1$ for $h_1$. Nevertheless a rough upper on $\alpha$ is required to be known in order to select the parameter $\gamma$ in Theorem 2.5.

**Theorem 2.5.** *We impose the conditions of Theorem 2.4; and we choose* $K = 1_{[0,1]}$*,* $s_n = n^\gamma$ *for some* $\gamma \in (0, 1/(1+2\alpha))$*; and* $h_4 = 1/s_n$*. Then there exist some positive constants* $b_0$ *and* $b_1$ *such that the estimator* $\hat{h}_1$ *in (2.6.1) satisfies*

$$\lim_{n\to\infty} P\Big(n^{1/(4+2\alpha)} \cdot \hat{h}_1 \in [b_0, b_1]\Big) \;=\; 1\,.$$

**Remark.** Note that we estimate the parameter $\alpha$ under general nonparametric constraints (see Assumption (A.2.8)), leading to the empirical bandwidth $\hat{h}_1$ in (2.6.1). If more restrictive parametric assumptions are imposed on the distribution of $\Delta U$ then the parameter $\alpha$ could also be estimated e.g. by maximum likelihood methods.

## 2.7 Simulation

For an illustration of the estimator in the univariate case, remember the panel data model in (2.2.1). Within this class of models, we constructed two leading specifications: a second and a third-order polynomial in the sole regressor $X_{k,t}$.

$$\begin{aligned} Y_{k,t} &= \Phi\left(X_{k,t}, A_k\right) + U_{k,t} \qquad \text{where:} \\ \Phi\left(X_{k,t}, A_k\right) &= A_{0,k} + A_{1,k}X_{k,t} + A_{2,k}X_{k,t}^2 \qquad \text{Quadratic 1D Model} \\ \Phi\left(X_{k,t}, A_k\right) &= A_{0,k} + A_{1,k}X_{k,t} + A_{2,k}X_{k,t}^2 + A_{3,k}X_{k,t}^3 \qquad \text{Cubic 1D Model} \end{aligned}$$

where, for all $k = 1, \ldots, n$ and $t = 1, \ldots, T$:

- $A_{j,k} \sim \mathcal{N}(0, .5) \qquad \forall \;\; j \in \{0, 1, 2, 3\}$
- $X_{k,t} \sim .5 + e_x \qquad e_x \sim \mathcal{N}(0, .5)$

- $U_{k,t} \sim e_v \qquad e_v \sim \text{Laplace}(0, .1)$

Since this is a univariate case, we can simply nonparametrically estimate the distribution of the conditional characteristic functions by using our estimator from Equation (2.4.6).

We select a proper $\alpha$ to optimize our results, and determine the bandwidths in the following way: $h_1 = n^{-1/(4+\alpha)}, h_2 = 2h_1, h_3 = h_1, h_0 = h_1^2$, as suggested by Theorem 2.4. While these are the asymptotically most efficient bandwidths, there may be better bandwidths in practical application. The restrictions that the bandwidths must obey imply that $0 < h_1 < h_2 < \rho$ and $h_3 > 0$.

We will compute the values of $\mu$ and $\sigma$ to minimize the Euclidean distance between $\widehat{\phi}_Z(s, x)$ and the characteristic normal distribution.

$$\phi_{\Delta Z}(s, x) = \exp(i\mu s - \sigma^2 s^2/2)$$

### 2.7.1 Results in the Baseline Specification

The specifications outlined above have easily represented true values. These are given by:

$$Z_{k,t} := \frac{\partial \Phi}{\partial x}(x, A)|_{x=X_{k,t}} = A_{1,k} + 2A_{2,k}X_{k,t} \qquad \text{Quadratic 1D Model}$$
$$Z_{k,t} := \frac{\partial \Phi}{\partial x}(x, A)|_{x=X_{k,t}} = A_{1,k} + 2A_{2,k}X_{k,t} + 3A_{3,k}^2 X_{k,t}^2 \qquad \text{Cubic 1D Model}$$

To display the true model, we use an oracle kernel density estimator that uses the (in the real world unobserved) values of $Z_{k.t}$. Figures 2.2 and 2.3 show the results comparing our estimator to the true distribution estimated by such an oracle kernel density estimator.

Start out by considering Figure 2.2: The blue line in the left two graphs corresponds to the true mean, resp., standard deviation, of the conditional marginal effects. The left two graphs display moreover the estimated conditional means, resp. standard deviations, for each value of $x$, and the corresponding estimation uncertainty as given by bootstrap 95% confidence bands. As is evident, the estimated means track the true values very closely, while the standard deviations perform (expectedly) worse, yet still deliver a quite satisfactory fit.

On the right are two contour graphs showing first a contour plot of the true conditional density of the marginal treatment effects along with the conditional means, as estimated using again an oracle kernel density estimator, and secondly an estimate of the conditional densities estimated using our method. As before, our estimator for the density of marginal effects matches the true distribution of the marginal effect very closely.

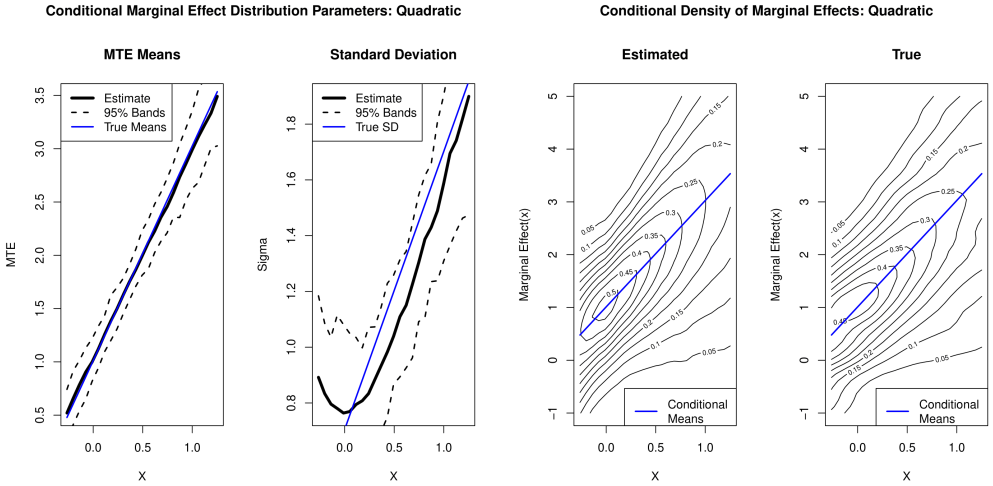

Figure 2.2: Estimates of quadratic 1D model using: $\alpha = 2$ and N = 10,000.The black line is our estimate. The dotted lines are our 95% confidence bands estimated with 100 bootstraps. The blue line is the true means and standard deviation we are trying to estimate.

Figure 2.3 then repeats the exercise for the cubic model and obtains similar, if slightly worse, performance, which is to expected given the slightly more complex model.

We also include an estimate of the quantiles of marginal effects in Figure 2.4, using our approach. This is done by inferring the quantiles from the conditional normal density, for which we have estimates of $\mu$ and $\sigma$ for each value of $X$.

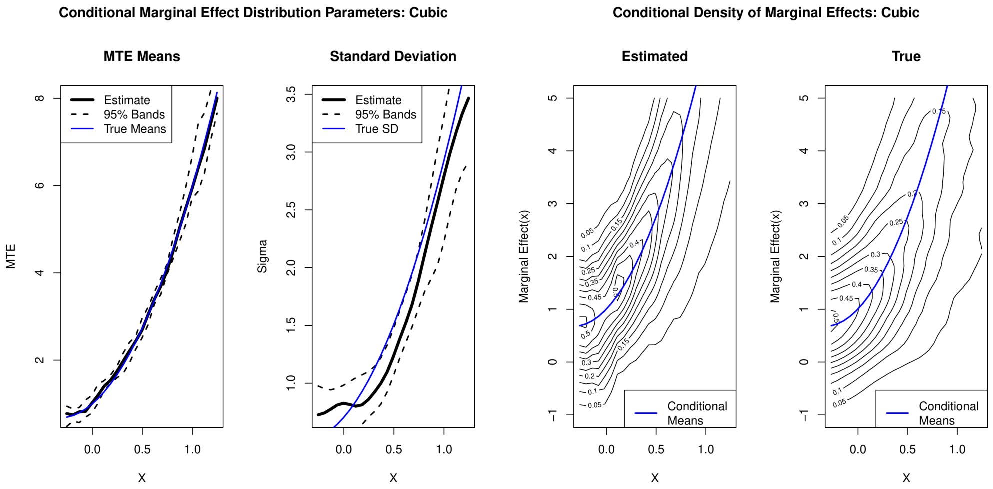

Figure 2.3: Estimates of cubic 1D model using: $\alpha = 2$ and N = 10,000. The dotted lines are our 95% confidence bands estimated with 100 bootstraps. The blue line is the true means and standard deviation we are trying to estimate.

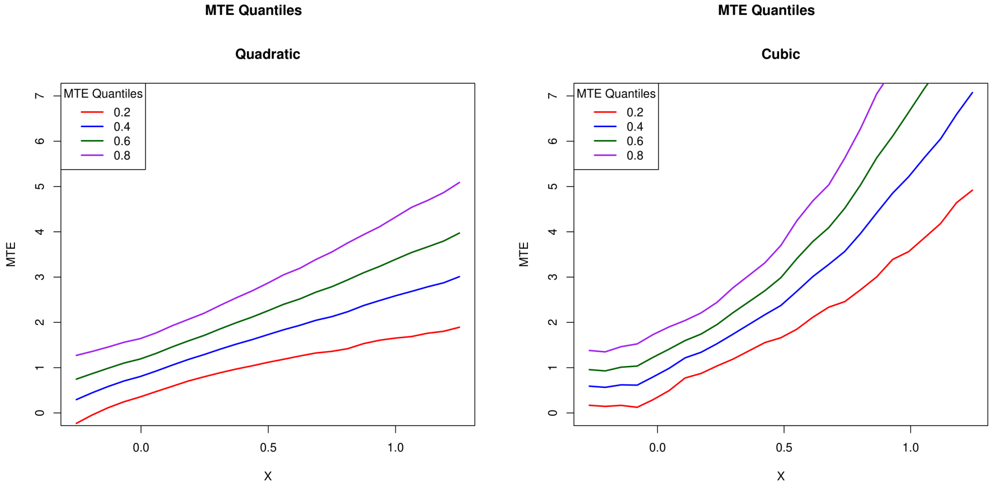

Figure 2.4: Estimates of Quantile Effects.

Note that these are conditional densities of marginal effects, so the most dense regions are on the boundaries where the standard deviation is the lowest, even though most of the data are near the mean of $X$. We can also estimate the joint densities of $Z := \frac{\partial \Phi}{\partial x}(x, A)$ and $X$, by multiplying our estimate of the conditional density with the density of $X$, $f(x)$. We estimate the density of $X$ using a kernel density estimation function. The resulting joint densities are displayed in Figure 2.5 below:

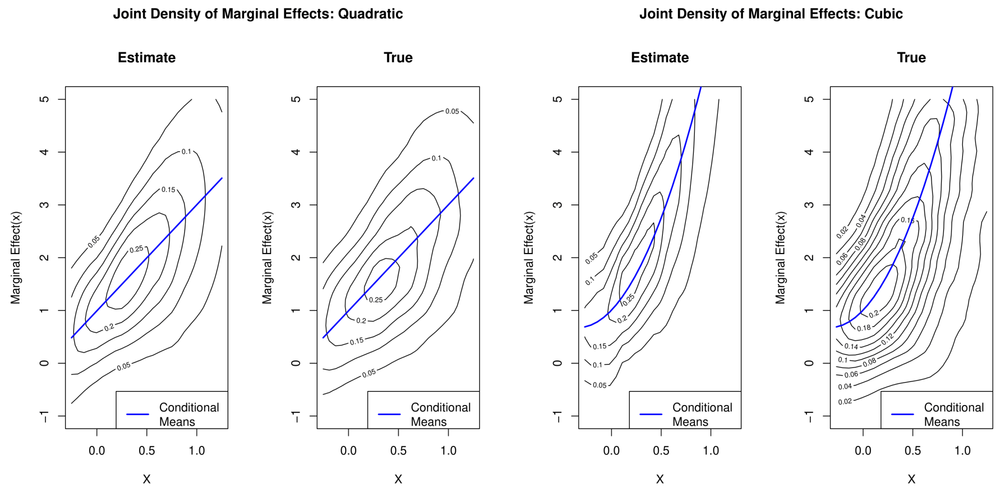

Figure 2.5: Estimates of joint distribution of the quadratic 1D model on the left and of the cubic 1D model on the right using the same parameters as above.

### 2.7.2 A Violation of Conditional Normality: Skewed Distribution of Effects

Next, in order to evaluate the robustness of our estimation procedure, we study the performance of our estimator in a simulation scenario which violates the conditional parametric assumption imposed for semiparametric estimation. We will assume that $A$ comes from a mixed normal distribution.

- $A_{j,k} \sim 0.5 \cdot \mathcal{N}(0.7, 0.2) + 0.5 \cdot \mathcal{N}(-0.25, 0.1) \qquad \forall \quad j \in \{0, 1, 2, 3\}$

This function is skewed to the right, i.e., it will not exhibit symmetrical marginal effects. The results for both the cubic case and quadratic case are included below. In Figures 2.6 and 2.7, we see that our estimates of the means are still quite accurate. However, our estimates for the standard deviation are slightly too high, since the estimated density exhibits a wider spread because of the skewed density of marginal effects.

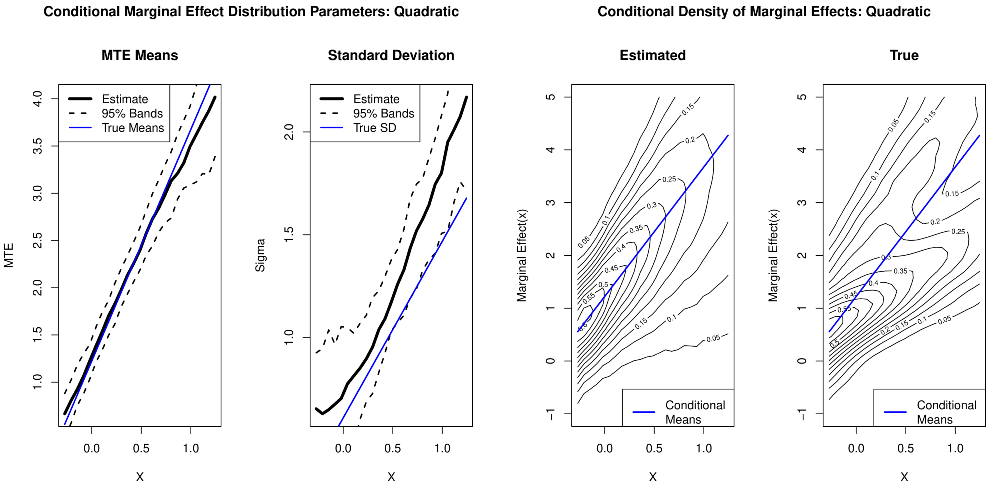

Figure 2.6: Estimates of quadratic 1D model using: $\alpha = 2$ and N = 10,000.The black line is our estimate. The dotted lines are our 95% confidence bands estimated with 100 bootstraps. The blue line is the true means and standard deviation we are trying to estimate.

Moreover, the joint and conditional estimated densities (see Figures 2.6, 2.7, and 2.8) do a reasonable job in capturing the general orientation of effects, but are unsurprisingly not fully able to capture the true model perfectly, as we (wrongly) impose normality of the conditional distribution. Note, however, that estimated conditional means are quite close to the true results, and the overall performance appears to be reasonably robust against violations of the parametric specification.

## 2.8 Empirical Application

In this section, we study the performance of our estimation procedure using real world data. We consider the estimation of the distribution of marginal effects of every additional dollar on the consumption of junk food. Because

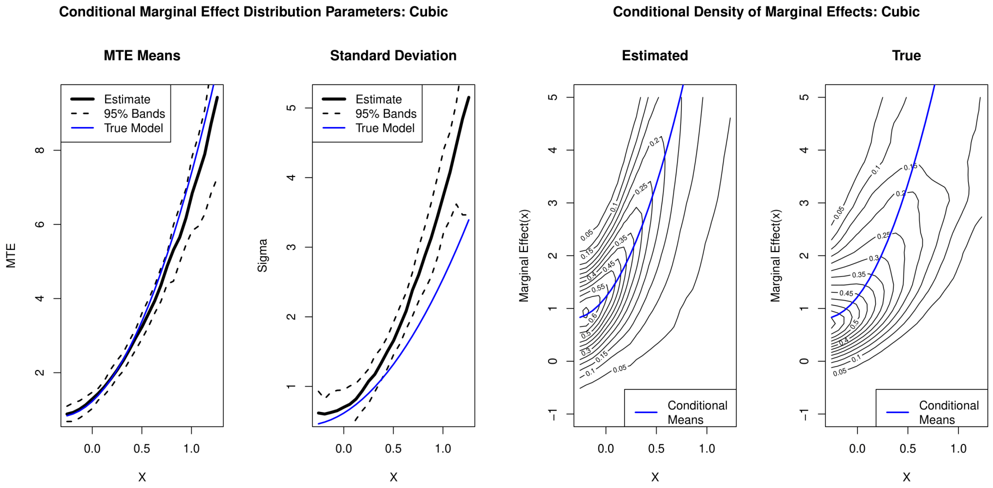

Figure 2.7: Estimates of cubic 1D model using: $\alpha$ = 2 and N = 10,000. The dotted lines are our 95% confidence bands estimated with 100 bootstraps. The blue line is the true means and standard deviation we are trying to estimate.

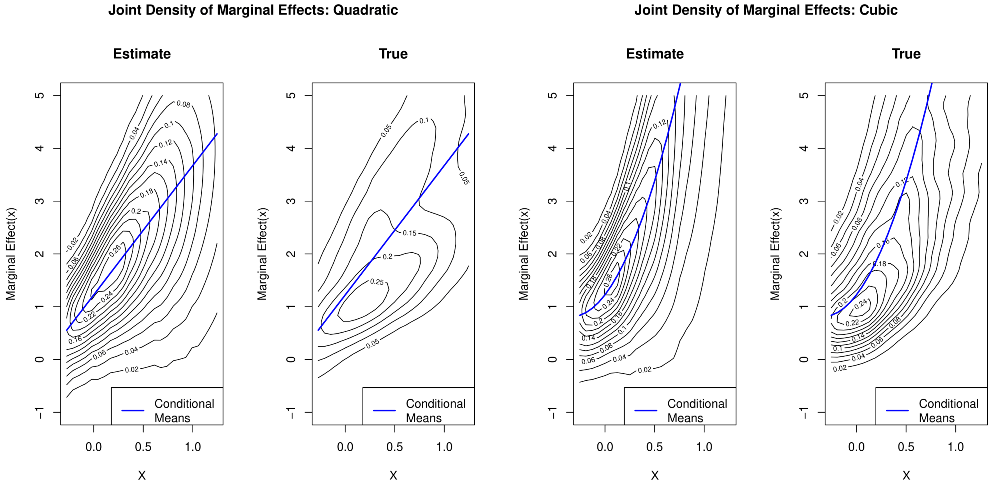

Figure 2.8: Estimates of joint distribution of the quadratic 1D model on the left and of the cubic 1D model on the right using the same parameters as above.

of the implied health consequences, as outlined below this question is highly policy relevant. In addition, our model is very well suited to capture differences in these marginal effects between wealthy and poor households, which are not captured at all by linear random coefficients models. This ability to exhibit differences for different wealth and income levels is crucial for the policy debate, as it is widely believed that excessive consumption of junk food is particularly prevalent at the lower end of the income distribution. As such, we hope that our estimator is able to inform this policy debate by providing a more nuanced picture of the distribution of marginal effects.

We start out with an overview of the data we use in our estimation exercise. After that, we provide a brief review of the policy debate surrounding junk food demand, especially with respect to differences in income. We then display our empirical findings which corroborate many of the suggestions put forward in the literature.

### 2.8.1 Data

#### 2.8.1.1 An Overview

For our application, we use the Nielsen Scanner Dataset which is available through the Kilts Center at the University of Chicago Booth School of Business[2]. We will focus our study on the year 2014 where there are about 55,000 individuals. This is a helpful dataset for estimating demand behavior since it contains detailed information based on price and quantity of all retail purchases as well as detailed household characteristics for all consumers. The data contain a representative sample of households in the United States who use in-home scanners to record all of their purchases intended for personal, in-home use.

[2] Researcher(s) own analyses calculated (or derived) based in part on data from The Nielsen Company (US), LLC and marketing databases provided through the Nielsen Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the Nielsen data are those of the researcher(s) and do not reflect the views of Nielsen. Nielsen is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

Nielsen matches the product scanned by the household to the actual price of the store where the product was bought. Nielsen estimates that about 30% of household consumption is accounted for by these purchases.

We will call this sum over all Nielsen expenditure categories total expenditure; under additive separability of the utility function this is the relevant total outlay variable. The same variable also takes the place in derivations involving economic rationality - under additive separability, this is the relevant "income" variable, e.g., to analyze Slutsky negative semidefiniteness. For this model, we estimate the total outlay ("income") and own price elasticities and the marginal effects of an additional unit of total outlay ("income") on the demand for junk food. Nielsen aggregates millions of universal product codes (UPC) into different groups of food.

We define junk food as any food classified as potato chips, candy or carbonated beverages by Nielsen. Junk food is a good example in our situation because these items lie on one extreme of the nutrition-taste trade-off (Blaylock et al., 1999). Junk food sacrifices almost all of its nutrition for taste. We aggregate the data to a monthly level such that period 1 is January 2014 and period 2 is February 2014. Of course, we could use different months as the time periods in our dataset as long as these periods exclude the irregular Christmas shopping period.

Prices are more precisily an aggregate price index called Stone-Lewbel (SL) cross section prices (see Lewbel (1989) and Hoderlein and Mihaleva (2008)). Generally speaking, SL prices use the fact that within a category of goods (junk food in our case), people have different tastes for the individual goods. Using standard aggregate price indices for junk food implicitly assumes that all individuals have identical Cobb Douglas preferences for all goods within this category, but SL prices allow all individuals to have heterogeneous Cobb Douglas preferences for the various commodities in this bundle. This implies that the typical approach of using aggregate price indices is a restrictive case of using SL prices. For this reason, SL prices should always be used when possible.

Total expenditure for all Nielsen goods and all junk food is aggregated each month as well. In order to get the proper expenditure, we only use households with two individuals and no children and divide expenditure by two, in order to estimate average expenditure per consumer. This is justified, as junk food is arguable a private good, and household composition effects can be expected to be negligible.

#### 2.8.1.2 Limitations

There are a few concerns with the data. The data rely on participants successfully recording their purchases in their home, so they may suffer from recording error. The specific issue that we might be concerned with is that consumers may consume a good when it is purchased and will not record the purchase when they return home. Einav et al. (2010) finds that consumable goods like soft drinks, chips, or candy are likely to be consumed before getting home so are more likely to not be scanned. There are also recording errors such as when a six-pack of goods are purchased and recorded as quantity six. However, these errors only seem to have minor effects. When compared to data from grocery store recorded sales, the data in Nielsen Homescan data matched 94% of the time (Einav et al., 2010).

Another potential source of measurement error is related to the price rather than the quantity. Individuals record their purchases by scanning the items they buy when they get home. The individuals input the quantity they purchase, and Nielsen matches it with the average price of the good at the store where they purchased it that week. This can lead to two types of errors. The first comes from the price changing in the middle of the week, though frequent changes during several weeks are less likely. The second type of error comes from not including discounts from loyalty cards. Einav et al. (2010) examines a retailer used in the Homescan data which has loyalty cards and finds that loyalty cards are used in about 75-80% of the transactions. Further, this would bias our prices and expenditure upwards. When comparing Homescan data with data from the retailer, Einav et al. (2010) finds that the prices used in the Homescan data is about 7% higher and the overall expenditure is 10% higher. On the other hand, these price measurement errors may be overestimated since some retailers do not have loyalty cards at all.

Finally, homescan data errors are comparable to errors found in other commonly used data sets. Aguiar and Hurst (2007) finds that life-cycle pattern of household expenditures recorded in Homescan Data is consistent with those reported for food expenditures at home in Panel Study of Income Dynamics (PSID). Einav et al. (2010) finds that these issues are not more serious than those in any other consumption surveys like the Current Population Survey (CPS). Lin (2018) compares the fraction of expenditures on different categories of products in the Nielsen Homescan Data and finds the results consistent to results from the Consumer Expenditure Survey (CES). In sum, we feel that these potential sources of measurement error may bias our results somewhat, but are unlikely to invalidate them.

### 2.8.2 Literature Review

There is a large literature on the determinants, extent and consequences of the consumption of junk food. As regards determinants, sometimes low-income propensity to consume unhealthy is attributed to the cost of healthy food (see, e.g.,Drewnowski and Darmon (2005), Golan et al. (2008), and Drewnowski and Eichelsdoerfer (2010)). However, Carlson and Frazão (2012) found that junk food is cheaper on a per-calorie basis than healthier foods like fruits, vegetables, whole grains and proteins, but that the healthier foods are actually cheaper on a per-serving basis. Rider et al. (2012) found that health attributes have been found to not be associated with higher average transaction prices.

When it comes to extent and possible consequences, obesity is one of the most important health problems in the United States, as well as many other countries. Many of the junk foods we consider are high in sugar, and excess sugar consumption is strongly linked with many diet-related diseases such as diabetes, cancers and heart disease (World Health Organization, 2015). Obesity leads to several hundred billion dollars spent on medical costs in the US annually, about 10-27 percent of all medical costs Finkelstein et al. (2009); Cawley et al. (2015). Thus, consumption of unhealthy food, such as junk food, can have a major impact on individual well being as well as the economy at large.

Our estimator allows for a more nuanced picture of the demand patterns for junk food, and hence enables policy makers to better target policy measures on subgroups of the population. Obesity and diabetes rates are higher for low income individuals (Drewnowski and Specter, 2004; Robbins et al., 2001). Binkley and Golub (2011) and Chen et al. (2012) all found that low-income households consume less nutritious foods. Allcott et al. (2017) showed that even when controlling for supply side factors, high-income households have a greater demand for healthy foods. We add to this literature a more differentiated description of the distribution of marginal effects for individuals with different incomes, which crucially relies on the added flexibility that our approach warrants relative to linear random coefficients models, e.g.,Graham and Powell (2012).

### 2.8.3 Income Elasticities and Marginal Effects of Income

To begin, as a building block for our model, but also to obtain naive “income” elasticities, we display the mean budget share of junk food (i.e., the proportion

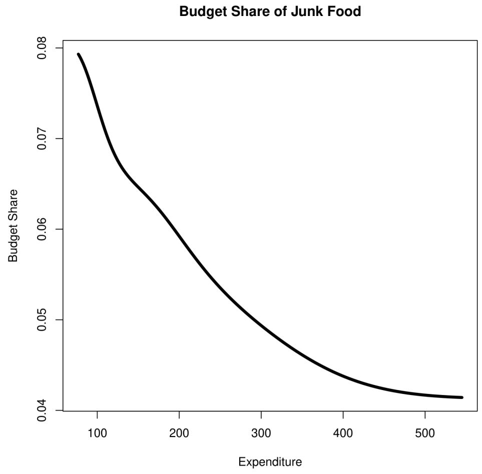

Figure 2.9: Nardaya-Watson kernel regression estimator of Budget Share of Junk Food based on total expenditures

of Nielsen recorded junk food over all Nielsen recorded items) for each household, $\omega_{k,t}$, as the dependent variable and total log expenditure, $E_{k,t}$, as the right hand variable in the first period (denoted $t$). Throughout this subsection, we control for prices by using households whose prices are in a neighborhood of the median price in period $t$, denoted $p$. Thus, the model we estimate is as follows:

$$\omega_{k,t} = \Phi\left(E_{k,t}, A_{k,t}, p_t\right) + U_{k,t} \tag{2.8.1}$$

The associated graph is included in Figure 2.9. Note that budget share is decreasing with total expenditure which strengthens the idea that low-income households eat more unhealthy food than high-income households. The convex curve implies that both the marginal effect of income on consumption of junk food and the income-elasticity of demand of junk food varies across expenditure. We will use our method to estimate $Z_j(e,p) = \frac{\partial \Phi}{\partial e} e$. We then follow standard arguments from Almost Ideal Demand System (AIDS) (Deaton, 1980), and use equation (2.8.2) estimate to identify and estimate the elasticity of income, $\varepsilon^d$

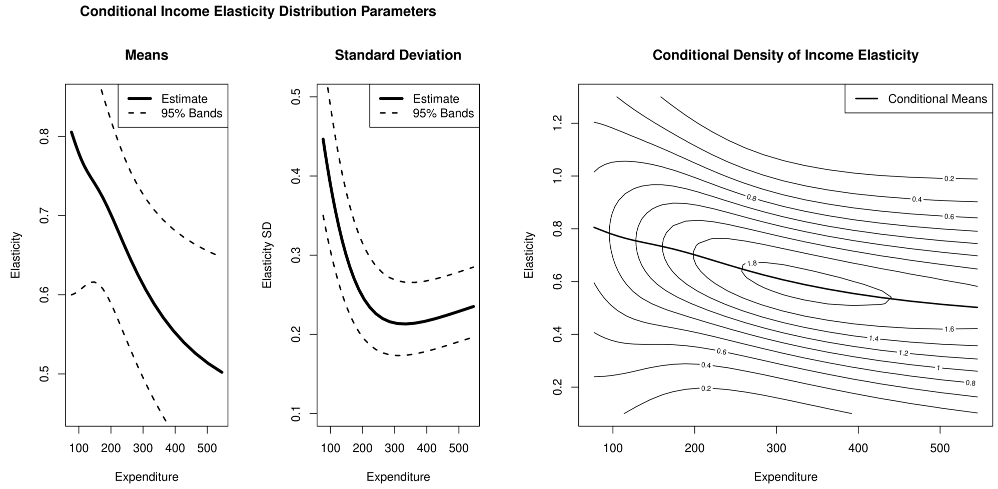

Figure 2.10: Estimates of Elasticity of demand using: $\alpha = 6$. For this sample, $N = 6,870$

using our estimate of $Z_j(e, p)$ from equation (2.8.1).

$$\varepsilon_j^d(e, p) = \frac{Z_j(e, p)}{\omega_j(e, p)} + 1 \tag{2.8.2}$$

To utilize this for the estimation of the elasticities, we use $\omega_j(e, p)$ which, as mentioned, is estimated using Nadaraya-Watson kernel regression estimator. This allows us then to estimate the conditional density of income elasticities of demand for junk food. The means and standard deviations of the coefficients, as well as the conditional density of marginal effects, are displayed in Figure 2.10. The pointwise standard errors have been constructed using the naive bootstrap. Note that the income elasticities of demand decrease with expenditure, and are clearly significantly non-linear. Thus, given an one percent increase in income, low-income individuals will increase their junk food consumption by a higher percentage than high-income individuals.

Note that these are estimates of the conditional density of income elasticities of the demand for junk food conditioned on "income" (as discussed, actually total Nielsen goods expenditure). We can estimate the joint density by multiplying this conditional density by the distribution of total expenditure, measured using a kernel density estimation. The result of this procedure can be seen in Figure 2.11 where we also include estimates of the conditional quantiles of the distribution of income elasticities and income.

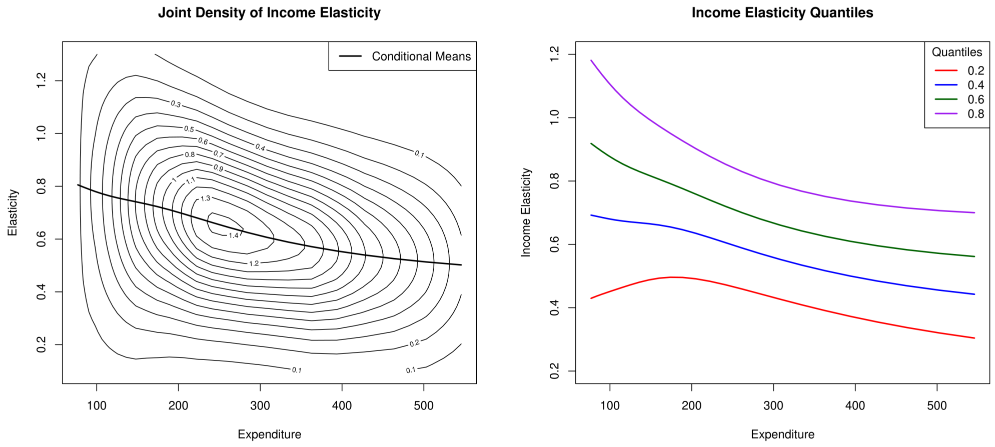

Figure 2.11: Joint distributions are calculated by multiplying the conditional distribution by the distribution of expenditure.

Furthermore, we can then use the elasticity estimates to estimate the density of marginal effects of an unit of additional income on the demand for junk food, using the following identity: Let $q$ be the quantity of junk food consumed. Consider

$$\varepsilon_d(e,p) = \frac{\partial \log(q)}{\partial \log(e)} = \frac{\partial q}{\partial e}\frac{e}{q} = \frac{\partial q}{\partial e}\frac{p}{\omega(e,p)} \tag{2.8.3}$$

Since we control for own price and keep it constant, we can normalize price to be equal to one for computational ease. Thus, we can estimate the marginal effect of an additional dollar on consumption of junk food, $\frac{\partial q}{\partial e}$. The result of this analysis is displayed in Figure 2.12, along with the quantile of these marginal effects. The effects follow the same trend as the income elasticities of demand, but the difference between low-income individuals and high-income individuals is more pronounced.

To understand this graph better we show, in Figure 2.13, the estimated density of marginal effects of income on consumption of junk food for different

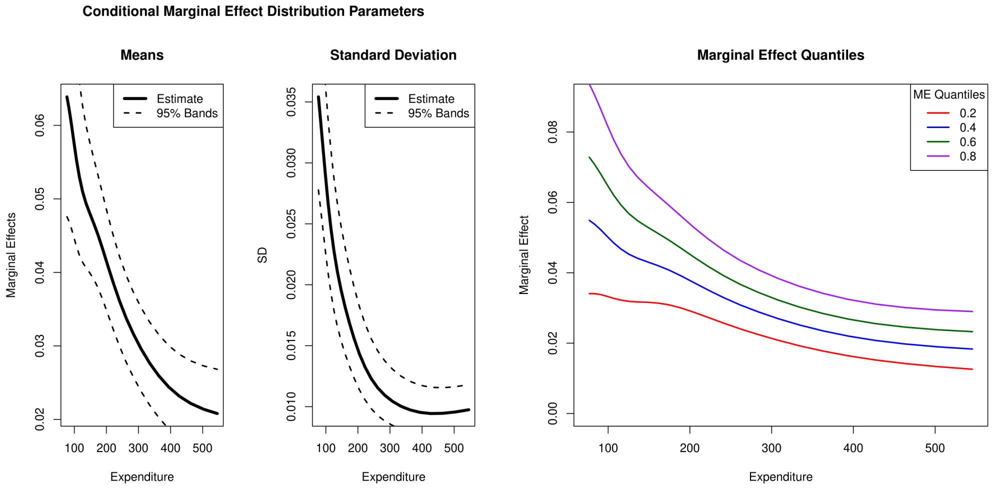

Figure 2.12: Estimates of the marginal effect of an additional dollar of expenditure on junk food using: $\alpha = 6$. For this sample, $N = 6,870$

groups based on their income quantile. Specifically, we graph the distribution of marginal effects for those at the .2, .4, .6 and .8 quantiles of the income distribution. To illustrate this point, consider the following example. In our example, low income individuals have income elasticities of about 0.8 and high income individuals have income elasticities of about 0.5. Consider that low income budget share of junk food is 0.08 while high income budget share of junk food is about 0.04. If we plug these values into equation (2.12), for low income individuals we obtain $0.8 = \frac{\partial q}{\partial e}\frac{1}{0.08}$ so that the marginal effect is $\frac{\partial q}{\partial e} \cong 0.064$. For high income individuals, $\frac{\partial q}{\partial e} \cong 0.02$. Thus, while the income elasticity of low income individuals is on average only 50% higher than the elasticity of high income individuals, the marginal effect of income on quantity of junk food consumed of poor individuals is more than twice as high compared to their high income counterparts. In other words, for every dollar they spend on Nielsen goods, they consume more than twice the quantity of junk food.

Remember that these densities of marginal effects are conditional on total expenditure ("income"). To estimate the joint density, as before we multiply the estimate of the conditional density by a kernel density estimate of total expenditure ("income"). The results for the joint density of marginal effects

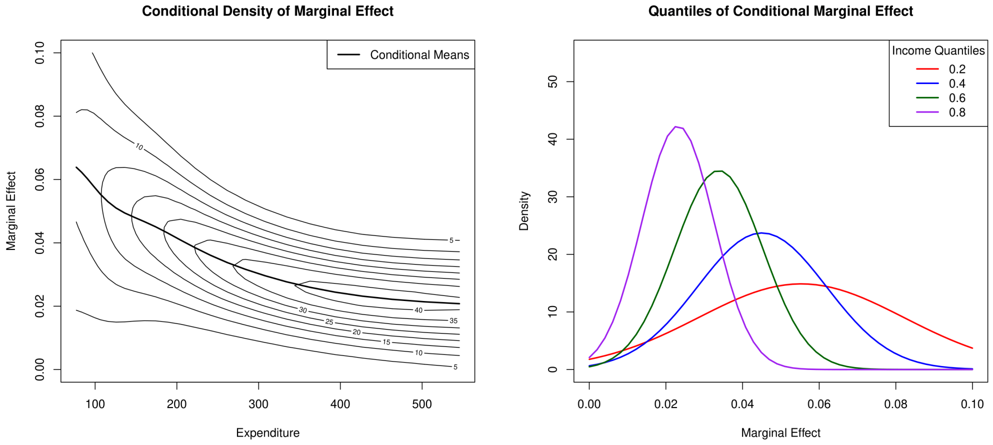

Figure 2.13: Conditional density and different expenditure quantiles of the estimates of marginal effect.

are found in Figure 2.14, along with the density of marginal effects for those in the .2, .4, .6 and .8 quantiles of the "income" distribution. As is to be expected, this reweighting results in the 0.6 quantile of the income distribution to deliver the density with largest values, rather than the edge case of the 0.8 quantile as is the case with the conditional densities.

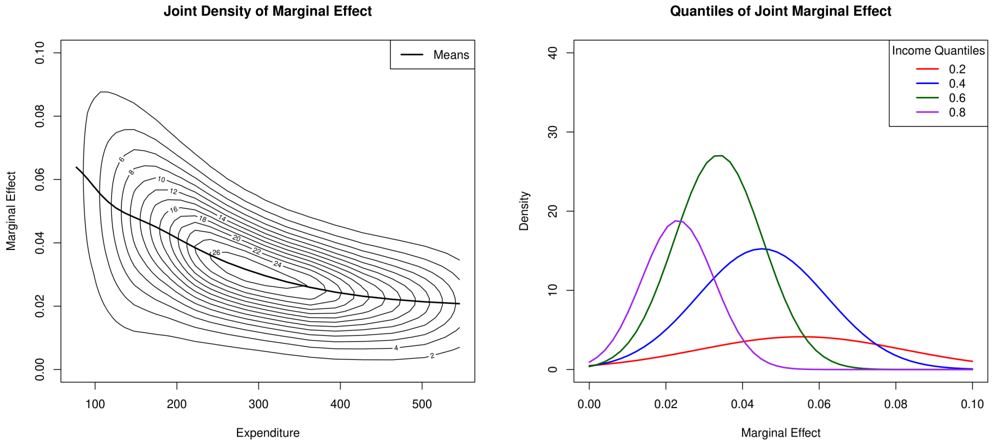

Figure 2.14: Joint distributions are calculated by multiplying the conditional distribution by the distribution of expenditure

Finally, note that a naive estimator could be based on an estimated derivative of the budget share graph in Figure 2.9. However, we expect these estimates to be biased because they do not account for the endogeneity stemming from the correlation between the high dimensional unobservables and income. The results are included below in Figure 2.15, which exhibit significant differences from our previous estimates.

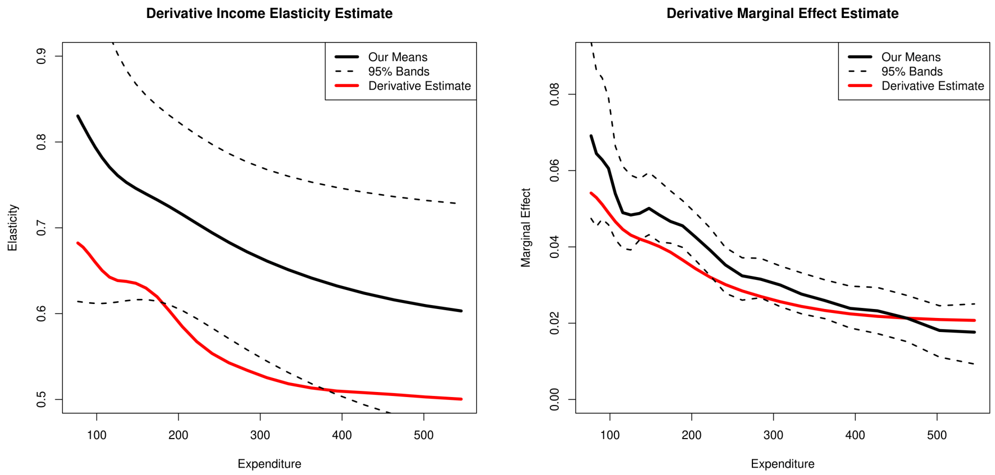

Figure 2.15: Mean and 95% bands of the mean of our estimates of income elasticity and marginal effect estimates compared to a nonparametric estimate of the derivative of the budget share graph.

Additional results with a different method to control for prices can be found in the appendix.

### 2.8.4 Own Price Elasticities

Following similar steps as above, we estimate own-price elasticities by using the budget share of junk food for each household, $\omega_{k,t}$, as the dependent variable and log of our SL price indices, $P_{k,t}$, as the right hand variable, but control again for income by selecting households with total expenditure close to the median, denoted $e$. Thus,

$$\omega_{k,t} = \Phi\left(e, P_{k,t}, A_{k,t}\right) + U_{k,t} \tag{2.8.4}$$

We will use our method to estimate $\tilde{Z}_j(p,e) = \frac{\partial\Phi}{\partial p}\left(p,e,A\right)|_{p=P_j}$, and use equation (2.8.5) to identify the elasticity of income, $\varepsilon^p$ using our estimate of $\tilde{Z}_j(e,p)$ from equation (2.8.4).

$$\varepsilon_j^p(e,p) = \frac{\tilde{Z}_j(e,p)}{\omega_j(e,p)} \tag{2.8.5}$$

We use again the Nadaraya-Watson estimator of $\omega_j(e,p)$, now as a function of price, see Figure 2.16.

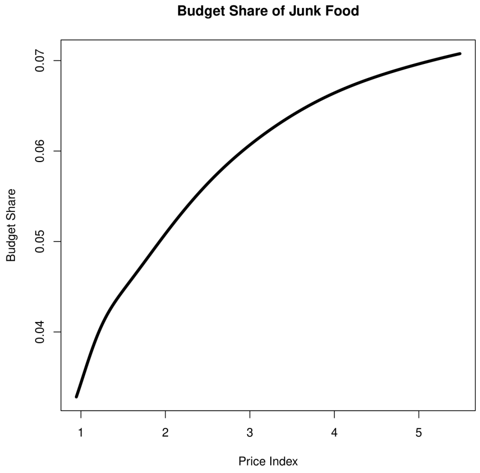

Figure 2.16: Nadaraya-Watson kernel regression estimator of Budget Share of Junk Food based on prices

With the estimate of budget share conditional on price, we can use our estimate of the density of $\tilde{Z}_j(e,p)$ and equation (2.8.5) to estimate the conditional distribution of own-price elasticities of for junk food. Below are the means and standard deviations of the coefficients as well as a contour map of the density in Figure 2.17, along with bootstrap standard errors. Note that own-price elasticities generally are negative and decrease with prices, i.e., increase in absolute value. Thus, given an increase of one percent in price, the reduction in demand for high-priced junk food is larger than for low-priced junk food.

Note again that these estimates are for the own-price elasticity for junk food conditional on price (and income). We can estimate the joint distribution by multiplying this conditional distribution by the density of expenditure, estimated using a kernel density estimation, see Figure 2.18 for the result. We also include the quantile estimates of own-price elasticities which allows to assess the difference in quantiles of consumers' own-price elasticities at different

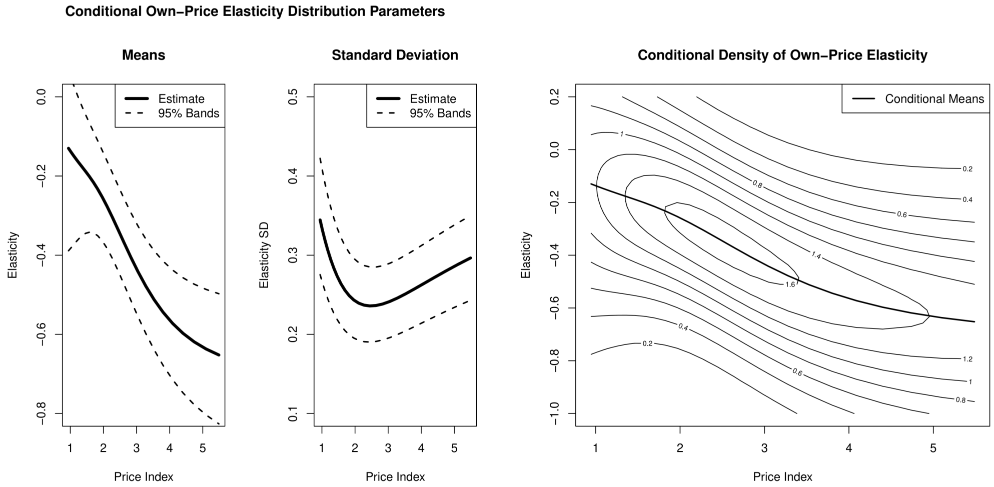

Figure 2.17: Estimates of Elasticity of demand using: $\alpha = 6$. For this sample, $N = 8,086$

prices.

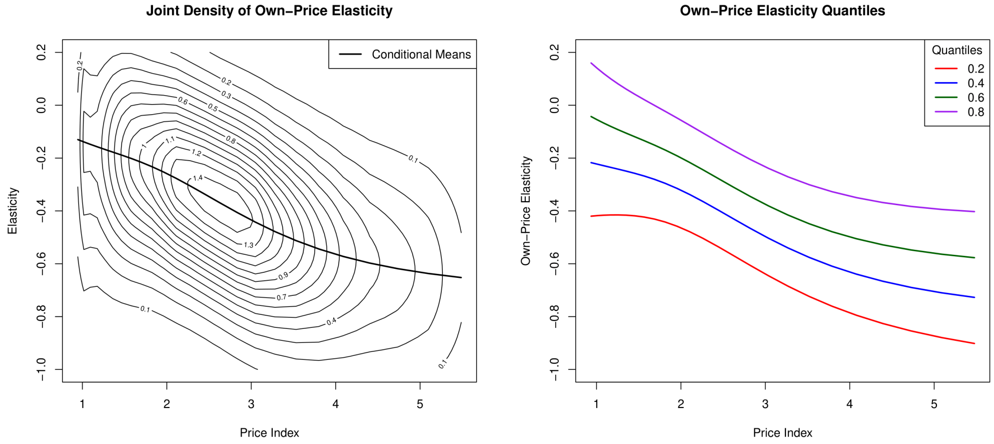

Figure 2.18: Joint distributions are calculated by multiplying the conditional distribution by the distribution of prices.

Finally, we compare our results again with the naive procedure that takes the derivative of the budget share regression, which differ because they do not properly account for the correlation stemming from the high dimensional correlated unobservables, see Fig 2.19.

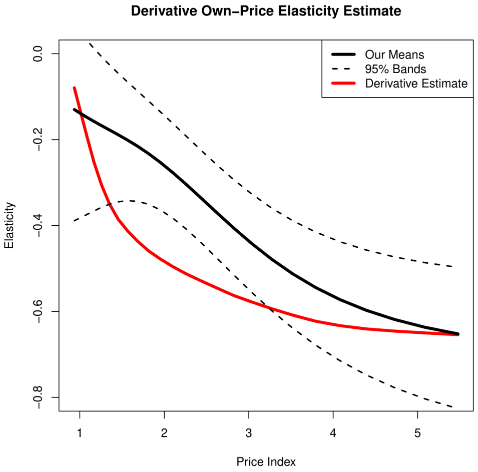

Figure 2.19: Mean and 95% bands of the mean of our estimates of own-price elasticity compared to a nonparametric estimate of the derivative of the budget share graph.

## 2.9 Appendix

### 2.9.1 Proofs

*Proof of Lemma 2.2.1*: As $\mathfrak{B}(\mathbb{R})$ is generated by a countable system of sets (e.g. consider the intervals $(-\infty, q]$, $q \in \mathbb{Q}$) the uniqueness theorem for probability measures guarantees that the measures $\zeta_j$ and $\tilde{\zeta}_j$ coincide almost surely by the assumptions of the lemma. Thus the set

$$\mathcal{Z}_j \;:=\; \big\{x \in \mathbb{R}^T \,:\, \zeta_j(x) \neq \tilde{\zeta}_j(x)\big\}\,,$$

is a $\mathcal{L}(X)$-null set; and $\mathcal{Z}_j$ is open in $\mathbb{R}^T$ thanks to the continuity of $\zeta_j$ and $\tilde{\zeta}_j$. Hence, the random vector $X$ lies in the closed set $\mathcal{S}_X \backslash \mathcal{Z}_j$ almost surely. As $\mathcal{S}_X$ is defined as the intersection of all those closed sets in which $X$ is located almost surely, it follows that

$$\mathcal{S}_X \;=\; \mathcal{S}_X \backslash \mathcal{Z}_j\,,$$

so that $\zeta_j(x) = \tilde{\zeta}_j(x)$ for all $x \in \mathcal{S}_X$. □

*Proof of Lemma 2.3.1*: For any $x \in \mathbb{R}$, we consider the $(T+1) \times (T+1)$-Vandermonde matrix $M(x)$ which contains $p(x_1)^\dagger, \ldots, p(x_T)^\dagger, p(x)^\dagger$ as its rows; and the matrix $N(x)$ which is obtained from $M(x)$ by replacing its last row by $q(x)$. Note that $\det N(x_j) = 0$ is equivalent to linear independence of the vectors $p(x_1), \ldots, p(x_T), q(x_j)$. Thanks to the multilinearity of the determinant and the well-known representation of determinants of Vandermonde matrices we deduce that

$$\det N(x) \;=\; \frac{d}{dx}\{\det M(x)\} \;=\; \Big(\prod_{1 \leq k < l \leq T} (x_l - x_k)\Big) \cdot \frac{d}{dx} \prod_{t=1}^{T} (x - x_t)\,.$$

Thus, $\det N(x_j)$ vanishes if and only if at least two of the $x_1, \ldots, x_T$ coincide or the polynomial $x \mapsto \prod_{t=1}^{T}(x - x_t)$ has a multiple zero at $x_j$. The latter claim requires at least one of the $x_t$ for $t \neq j$ to coincide with $x_j$, which implies the first claim. □

*Proof of Lemma 2.3.2*: We easily recognize by definition that the vectors $p(x_1), \ldots, p(x_T), q(x_j)$ are all continuous functions in $x \in \mathbb{R}^T$. Applying a Gram-Schmidt process we obtain that

$$\begin{aligned}
p_k^*(x) &= p(x_k) - \sum_{l=1}^{k-1} \big(p(x_k)^\dagger p_l^*(x)\big)\, p_l^*(x) / \big|p_l^*(x)\big|^2, \quad k = 1, \ldots, T, \\
q_j^*(x) &= q(x_j) - \sum_{l=1}^{T} \big(q(x_j)^\dagger p_l^*(x)\big)\, p_l^*(x) / \big|p_l^*(x)\big|^2, \\
\tau_j(x) &= \big|q_j^*(x)\big|^2,
\end{aligned}$$

for $x \in \mathcal{X}$ so that $\tau_j$ is continuous on $\mathcal{T}_X$ as well. The positivity of $\tau_j$ is an immediate consequence of Lemma 2.3.1 as $\tau_j(x) = 0$ implies linear dependence between $p(x_1), \ldots, p(x_T), q(x_j)$. □

*Proof of Lemma 2.3.3*: For any $x, y \in \mathbb{R}^T$, $b \geq 0$, we deduce by the triangle inequality that

$$\begin{aligned}
\mathcal{F}\big(\zeta_j^{[b]}(x), \zeta_j^{[b]}(y)\big) \;\leq\; & \mathcal{F}\big(\mathcal{L}(A^{[0]} \mid X = x), \mathcal{L}(A^{[0]} \mid X = y)\big) \\
& + \sup_{s \in \mathbb{R}} \big|\psi_{A^{[0]}|X=x}\big(sq(x_j)\big) - \psi_{A^{[0]}|X=x}\big(sq(y_j)\big)\big| \\
& + \sup_{s \in \mathbb{R}} \big|\psi_{A^{[0]}|X=x}\big(sq(x_j)\big)\big| \cdot \Big|\exp\Big(-\frac{1}{2}bs^2\tau_j^2(x)\Big) - \exp\Big(-\frac{1}{2}bs^2\tau_j^2(y)\Big)\Big|.
\end{aligned} \tag{2.9.1}$$

The first term in (2.9.1) converges to 0 as $y \to x$ by Assumption (A.2.4). As $A^{[0]}$ has a conditional Lebesgue density given $X = x$ it follows from the Riemann-

Lebesgue lemma (see e.g. Bochner et al. (1949)) that $\lim_{|u|\to\infty} \psi_{A^{[0]}|X=x}(u) = 0$. Thus, for any $\varepsilon > 0$, there exists some $R > 0$ such that $\big|\psi_{A^{[0]}|X=x}(u)\big| < \varepsilon/4$ for all $u$ with $|u| > R$. Since $|q(x)| \geq 1$ for all $x \in \mathbb{R}$ the second term in (2.9.1) obeys the upper bound

$$\varepsilon/2 + \sup_{|s|\leq R} \big|\psi_{A^{[0]}|X=x}\big(sq(x_j)\big) - \psi_{A^{[0]}|X=x}\big(sq(y_j)\big)\big| \,. \tag{2.9.2}$$

As the function $x \mapsto q(x)$ is continuous and any characteristic function is uniformly continuous, (2.9.2) is bounded from above by $\varepsilon$ whenever $|y - x|$ is sufficiently small with respect to only $\varepsilon$ and $R$. Therefore the second term tends to 0 as $y \to x$.

It remains to consider the third term in (2.9.1). Let $\varepsilon$ and $R$ be as in the previous paragraph. Then the third term is smaller or equal to

$$\varepsilon/2 + \sup_{|s|\leq R} \Big|\exp\Big(-\frac{1}{2}bs^2\tau_j^2(x)\Big) - \exp\Big(-\frac{1}{2}bs^2\tau_j^2(y)\Big)\Big| \,. \tag{2.9.3}$$

As $x \mapsto \tau_j(x)$ is continuous (see Lemma 2.3.2) and the exponential mapping is uniformly continuous on any bounded domain, the term (2.9.3) is bounded from above by $\varepsilon$ whenever $|y-x|$ is sufficiently small with respect to $\varepsilon$ and $R$. Finally we have shown that all three terms in (2.9.1) converge to 0 as $y$ tends to $x$. □

*Proof of Lemma 2.3.4*: Applying Fourier transformation to both sides of the given equality we obtain that

$$Q^{ft}(x) \cdot \exp\Big(-\frac{1}{2}\alpha x^2\Big) = Q^{ft}(x) \cdot \exp\Big(-\frac{1}{2}\alpha'^2\Big), \quad \forall x \in \mathbb{R}\,.$$

As $Q^{ft}$ is continuous and satisfies $Q^{ft}(0) = 1$ there exists a non-void open neighborhood of 0 in which $Q^{ft}$ does not vanish. Therefore the functions

$x \mapsto \exp\big(-\alpha x^2/2\big)$ and $x \mapsto \exp\big(-\alpha'^2/2\big)$ coincide on this neighborhood so that $\alpha = \alpha'$. □

*Proof of Theorem 2.3*: Thanks to (2.5.4) and the compact support of $f_X$, which is guaranteed by Assumption (A.2.5'), we may easily verify the first part of Assumption (A.2.6) for some $c_\Phi$ sufficiently large. With respect to the second part we deduce that

$$\mathcal{F}\big(\zeta_1(y), \zeta_1(z)\big) \;\le\; c_\zeta \cdot |y - z| \,,$$

for all $y, z \in \mathbb{R}$ where $c_\zeta := \|K'\|_\infty/2$. Thus Assumption (A.2.6) holds true.

As the statistic $\Delta Y_j$, $j = 1, \ldots, n$, has been shown to be sufficient for $\zeta_1(x)$ and, hence, for the parameter $\theta$, we may consider $P_\theta^{(n)}$ as the image measure of this statistic. Now we put $\theta_n := 3d \cdot n^{-1/(4+2\alpha)}$ so that at least one of the events $\{|\hat{\theta}_n - \theta_n| > d \cdot n^{-1/(4+2\alpha)}\}$ and $\{|\hat{\theta}_n| > d \cdot n^{-1/(4+2\alpha)}\}$ occurs. For sufficiently large $n$ it holds that

$$\sup_{\theta \in [0,1]} P_\theta^{(n)}\big[|\hat{\theta}_n - \theta| \;>\; d \cdot n^{-1/(4+2\alpha)}\big] \;\ge\; \frac{1}{2} - \frac{1}{2}\,\mathrm{TV}\big(P_{\theta_n}^{(n)}, P_0^{(n)}\big)\,.$$

By standard information-theoretic arguments, we deduce that

$$\begin{aligned}&\mathrm{TV}\big(P_{\theta_n}^{(n)}, P_0^{(n)}\big) \;\le\\ &2\Big\{\Big(1 + E\chi^2\big(f_{B|X}^{(\theta_n)} * f_{\Delta U}(\cdot/(X_1 - X_2)), f_{B|X}^{(0)} * f_{\Delta U}(\cdot/(X_1 - X_2))\big)\Big)^n - 1\Big\}^{1/2}\end{aligned}$$

where $\chi^2$ stands for the $\chi^2$-distance between two measures. By Parseval's identity, it holds that

$$
\begin{aligned}
& E\chi^2\big(f_{B|X}^{(\theta_n)} * f_{\Delta U}(\cdot/(X_1-X_2)), f_{B|X}^{(0)} * f_{\Delta U}(\cdot/(X_1-X_2))\big) \\
& \quad \le \text{ const.} \cdot \theta_n^2 \cdot E\,K^2\big(|X-x|/\theta_n\big) \\
& \qquad\qquad \cdot \int \big|\{f_0 \cos(4\cdot)\} * f_{\Delta U}(\cdot/(X_1-X_2))\big|^2(t)(1+t^4)dt \\
& \quad = \text{ const.} \cdot \theta_n^2 \cdot \max\Big\{E\,K^2\big(|X-x|/\theta_n\big)\,|X_1-X_2|^{-2\ell_2} \\
& \qquad\qquad \cdot \int \big|\{f_0^{ft}\}^{(\ell_1)}(t\pm 4)\big|^2\,\big|\psi_{\Delta U}^{(\ell_2)}(t/(X_1-X_2))\big|^2 \\
& \qquad\qquad : \ \ell_1,\ell_2 \in \mathbb{N}_0, \ell_1+\ell_2 \le 2\Big\} \\
& \quad = \mathcal{O}\big(\theta_n^{4+2\alpha}\big)\,.
\end{aligned}
$$

Therefore, choosing $d > 0$ sufficiently small, we may ensure that

$$
\limsup_{n\to\infty} \mathrm{TV}\big(P_{\theta_n}^{(n)}, P_0^{(n)}\big) \ < \ 1\,,
$$

which completes the proof of the theorem. □

*Proof of Theorem 2.4*: Writing

$$
\begin{aligned}
N_0 &:= \sum_{k=1}^n 1_{\mathcal{S}_X^{(h_0)}}(X_{k,1}, X_{k,2})\,, \\
N_1 &:= \sum_{k=1}^n 1_{\mathcal{S}_X^{(h_1,h_2)}}(X_{k,1}, X_{k,2}) \cdot 1_{[0,h_3]}\big(|X_{k,1}-x_1|\big)\,,
\end{aligned}
$$

we introduce the events

$$
\begin{aligned}
\mathcal{E}_0 &:= \{N_0 \ge c\cdot nh_0\}\,, \\
\mathcal{E}_1 &:= \{N_1 \ge c\cdot nh_3(h_2-h_1)\}\,,
\end{aligned}
$$

for some constant $c > 0$. By Chebyshev's inequality and Assumption (A.2.5') we deduce that the probabilities for the complements of $\mathcal{E}_0$ and $\mathcal{E}_1$ converge to zero as $n$ tends to infinity for $c > 0$ sufficiently small. The events $\mathcal{E}_0$ and $\mathcal{E}_1$ are contained in the $\sigma$-field $\sigma_X$ which is generated by the random variables $X_{k,t}$, $k = 1, \ldots, n$, $t = 1, 2$.

Now put $\varepsilon_n := dn^{-1/(4+2\alpha)}$ for some constant $d > 0$. By Assumption (A.2.9) the inequality

$$\int_{-R}^{R} \big|\{\zeta_1(\hat{\theta}; x)\}^{ft}(s) - \{\zeta_1(\theta; x)\}^{ft}(s)\big|^2 ds \;\geq\; c_p^2\, \varepsilon_n^2\,,$$

holds true on the event $\{|\hat{\theta} - \theta| > \varepsilon_n\}$. Then it follows from the definition of $\hat{\theta}$ that

$$\int_{-R}^{R} \big|\hat{\psi}_{Z_1}^{(h_0,h_1,h_2,h_3)}(x; s) - \{\zeta_1(\theta; x)\}^{ft}(s)\big|^2 ds \;\geq\; \frac{1}{4} c_p^2\, \varepsilon_n^2\,,$$

whenever $|\hat{\theta} - \theta| > \varepsilon_n$. Hence, by Markov's inequality, we deduce that

$$\begin{aligned} P\big[|\hat{\theta} - \theta| > \varepsilon_n\big] \;\leq & \,4 c_p^{-2}\, \varepsilon_n^{-2} \\ & \cdot \int_{-R}^{R} E\, 1_{\mathcal{E}_0 \cap \mathcal{E}_1}\, \big|\hat{\psi}_{Z_1}^{(h_0,h_1,h_2,h_3)}(x; s) - \{\zeta_1(\theta; x)\}^{ft}(s)\big|^2 ds \\ & + \, 1 - P(\mathcal{E}_0 \cap \mathcal{E}_1)\,. \end{aligned} \tag{2.9.4}$$

By a standard bias-variance decomposition for the conditional expectation, the Cauchy-Schwarz inequality and Assumption (A.2.6), we obtain that

$$\begin{aligned} & E\big\{\big|\hat{\psi}_{Z_1}^{(h_0,h_1,h_2,h_3)}(x; s) - \{\zeta_1(\theta; x)\}^{ft}(s)\big|^2 \mid \sigma_X,\, \hat{\psi}_{\Delta U}^{(h_0)}\big\} \\ & \quad \leq\; (2\rho_n + 1)/\big\{\rho_n \,+\, \hat{\Xi}_U\big\} \,+\, 4\,\big\{c_\Phi R h_2/2 + c_\zeta(2h_3 + h_2)\big\}^2 \,+\, 4\,\hat{\Xi}_\Delta \,/\, \big\{\rho_n \,+\, \hat{\Xi}_U\big\}\,, \end{aligned} \tag{2.9.5}$$

for all $s \in [-R, R]$ where $\sigma_X$ denotes the $\sigma$-field generated by $X_1, \ldots, X_n$; and

$$\begin{aligned}
\hat{\Xi}_U \;&:=\; \sum_{k=1}^n \big|\hat{\psi}_{\Delta U}^{(h_0)}(s/\Delta X_k)\big|^2 \cdot 1_{\mathcal{S}_X^{(h_1,h_2)}}(X_{k,1}, X_{k,2}) \cdot 1_{[0,h_3]}(|X_{k,1} - x_1|)\,, \\
\hat{\Xi}_\Delta \;&:=\; \sum_{k=1}^n \big|\hat{\psi}_{\Delta U}^{(h_0)}(s/\Delta X_k) - \psi_{\Delta U}(s/\Delta X_k)\big|^2 \cdot 1_{\mathcal{S}_X^{(h_1,h_2)}}(X_{k,1}, X_{k,2}) \cdot 1_{[0,h_3]}(|X_{k,1} - x_1|)\,, \\
\Xi_U \;&:=\; \sum_{k=1}^n \big|\psi_{\Delta U}(s/\Delta X_k)\big|^2 \cdot 1_{\mathcal{S}_X^{(h_1,h_2)}}(X_{k,1}, X_{k,2}) \cdot 1_{[0,h_3]}(|X_{k,1} - x_1|)\,.
\end{aligned}$$

We deduce by Assumption (A.2.6) that

$$E\big(\hat{\Xi}_\Delta \mid \sigma_X\big) \;\leq\; N_1/N_0 \;+\; R^2 c_\Phi^2 \, \Xi_U \, h_0^2/h_1^2\,. \tag{2.9.6}$$

Thus, on the event $\mathfrak{E}_3(s) := \{\hat{\Xi}_U > \Xi_U/2\}$, $|s| \leq R$, the conditional expectation of term (2.9.5) given $\sigma_X$ obeys the upper bound

$$\mathcal{O}\big(h_2^2 + h_3^2 + h_0^2/h_1^2 + 1/\Xi_U + N_1/(\Xi_U N_0)\big)\,, \tag{2.9.7}$$

where $\Xi_u$ has the asymptotic lower bound $N_1 \cdot h_1^{2\alpha}$ with uniform constants by the Assumptions (A.2.5') and (A.2.8). On the complement of $\mathfrak{E}_3(s)$, the conditional expectation of term (2.9.5) given $\sigma_X$ is bounded from above by

$$\mathcal{O}(n^2) \cdot \exp\big\{ - N_0(1 - 1/\sqrt{2} - c_\Phi h_0/h_1)^2 c_{U,1}^2 (1 + R/h_1^\alpha)^{-2}/8\big\}\,, \tag{2.9.8}$$

by Assumption (A.2.6) and Hoeffding's inequality. Applying the expectation to the terms (2.9.7) and (2.9.8) – multiplied by $1_{\mathcal{E}_0 \cap \mathcal{E}_1}$ – we conclude that the right hand side of (2.9.4) tends to zero if, first, the limit superior is taken with respect to $n \to \infty$ and, then, the limit $d \to \infty$ is applied. □

*Proof of Theorem 2.5*: It suffices to show the existence of some $c > 0$ such that

$$\limsup_{n\to\infty} P\Big(\big|\hat{\alpha} - \alpha\big| \,>\, c/\log n\Big) \,=\, 0\,. \tag{2.9.9}$$

Using that the probability of $\mathcal{E}_4$ (equivalent to the event $\mathcal{E}_0$ from the proof of Theorem 2.4 when replacing $h_0$ by $h_4$) converges to 1; that (2.4.2) holds true; and Hoeffding's inequality – conditionally on $\sigma_X$ – we can verify (2.9.9) when $c$ is sufficiently large with respect to $\gamma$. □

### 2.9.2 Summary Statistics

Below is the Summary Statistics for the data we used in our empirical application.

Table 2.1

| | January 2014 | February 2014 |
|---|---|---|
| SL Price Index | 0.7751 | 0.8069 |
| | (0.5850) | (0.5769) |
| Junk Food Share | 0.0567 | 0.0639 |
| | (0.0596) | (0.0631) |
| Total Expenditure | 477.96 | 448.33 |
| | (325.43) | (302.09) |

This table contains the mean and standard deviation (in parenthesis beneath the means) for the variables that we use in our analysis

### 2.9.3 Application with Different Prices

Below are the results when we control for prices a little differently. Here, price is controlled such that price is centered around the .4 quantile. This serves as a robustness check on the results from our empirical application of the paper. The overall trends are consistent in both cases.

The only difference of significance is that the decline of mean Elasticity of Demand does not change as much for low-income vs. high-income individuals (see Figure 2.20 compared to Figure 2.10). For example, in our base case,

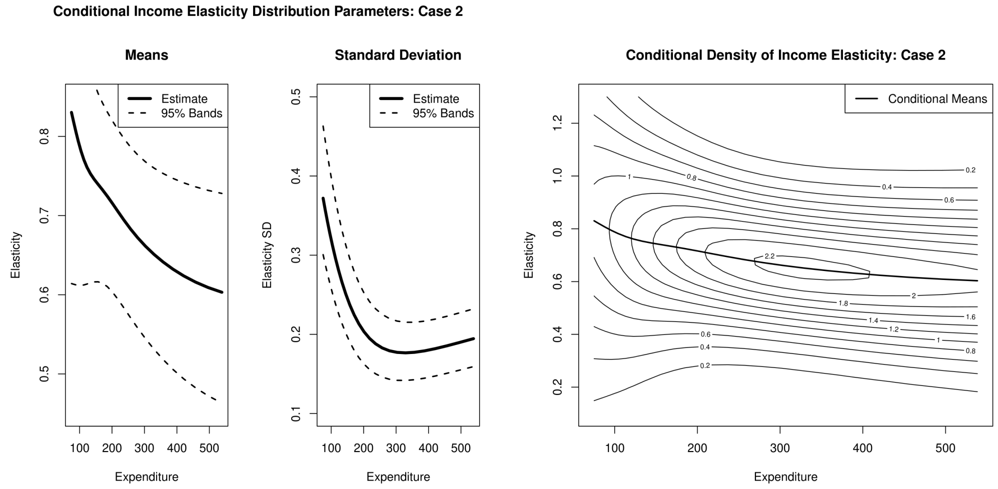

Figure 2.20: Estimates of Elasticity of demand using: $\alpha = 6$. For this sample, $N = 8,631$

mean income elasticity for low-income individuals is about 0.8 and for high income individuals it is about 0.5. In our adjusted case, the income elasticity of low-income individuals is 0.8 while for high income individuals it's about 0.6. This is a minor difference and the results from these estimates easily fit in our confidence bands from our paper.

Our marginal effects estimation in this case is also very similar (see Figures 2.12 and 2.22). These results imply that these results are consistent across different prices, as long as prices are properly controlled for.

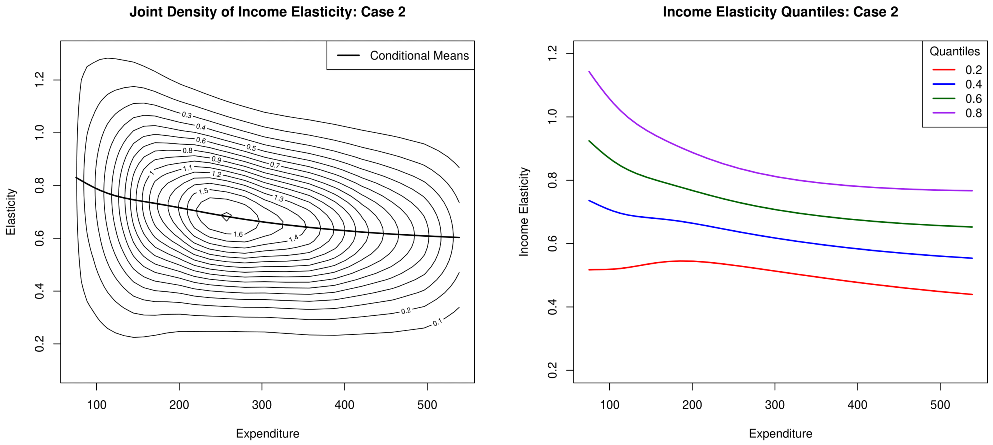

Figure 2.21: Joint distributions are calculated by multiplying the conditional distribution by the distribution of expenditure.

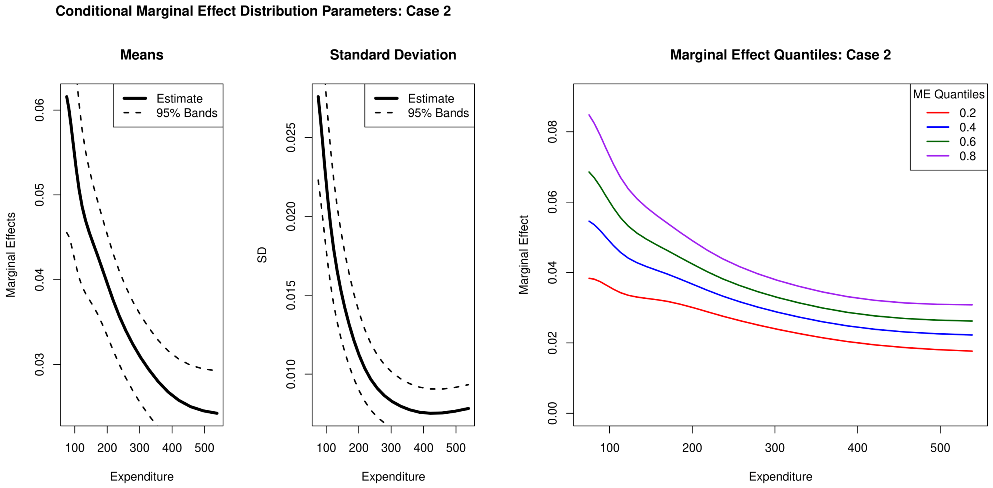

Figure 2.22: Estimates of the marginal effect of an additional dollar of expenditure on junk food using: $\alpha = 6$. For this sample, $N = 8,631$

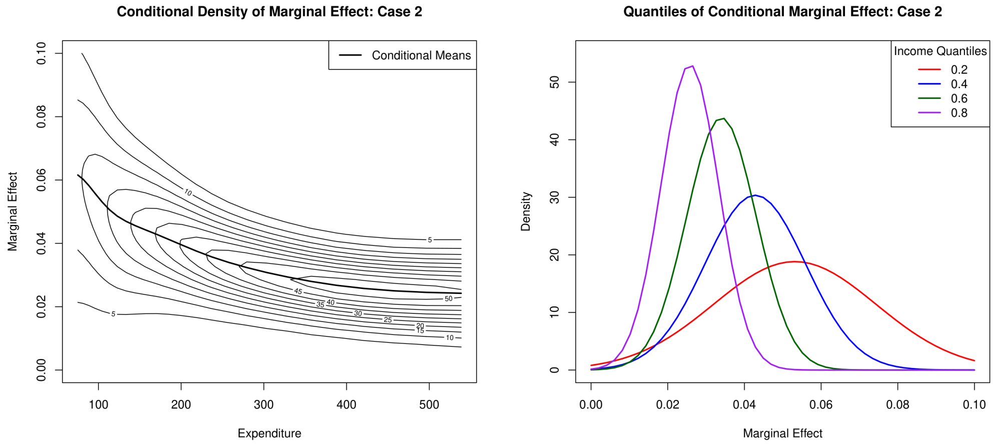

Figure 2.23: Conditional density and different expenditure quantiles of the estimates of marginal effect.

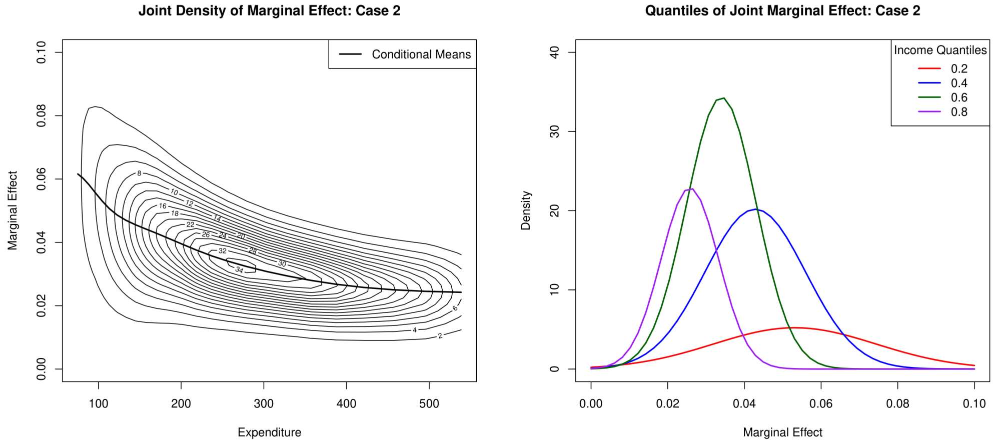

Figure 2.24: Joint distributions are calculated by multiplying the conditional distribution by the distribution of expenditure

# Chapter 3

# A Binary Choice Difference-in-Differences Model with Heterogeneous Treatment Effects and an Application on Soda Taxes

## 3.1 Introduction

Difference-in-Differences Models are a well established method for policy evaluation and treatment effect estimation in the applied microeconomics literature. Some early examples of the use of Difference-in-Differences include Ashenfelter (1978) and Ashenfelter and Card (1985), who have used Difference-in-Differences methods to quantify the effects of training programs on earnings. Other well-known examples are Card (1990) and Card and Krueger (1994), who have explored the effects of exogenous changes in labor market conditions on employment and wages. A more recent paper by Abadie and Dermisi (2008) estimates the effects of a terrorist threat on agglomeration economies in central business districts using a Difference-in-Differences approach.

The reasons for the abundant use of Difference-in-Differences models in applied research are manifold. First, Difference-in-Differences models are easy to implement, as they can be estimated via simple OLS. This allows for covariates to be included as additional regressors in a straightforward way. Second, one of the coefficients from the OLS regression can be directly interpreted as the Average Treatment Effect on the Treated (ATT) which is usually the object of interest in the treatment effect literature. Third, the Difference-in-Differences approach does not require a strict exogeneity assumption, as the two-period panel structure of the data allows us to deal with time-invariant unobserved heterogeneity.

Binary outcome variables are very common in applied economics research, e.g. when studying labor force participation decisions, retirement decisions, fertility decisions, etc. Unfortunately, the standard linear Difference-in-Differences model breaks down for binary dependent variables as soon as continuous covariates are included in the estimation procedure. The classical OLS assumptions (constant marginal effects, normality of the error term, and homoskedastic error terms) are then usually violated and thus standard inference will be wrong.

A possible remedy is to use Probit or Logit models which are able to account for the non-linearity inherent in models with binary outcome variables, but do not carry over some of the other convenient properties of standard Difference-in-Differences model. Namely, the standard Probit and Logit models do not allows for correlation between treatment status and time invariant unobservables. Also, the marginal probability of the interaction term of the pre-post treatment dummy and the treatment-control group dummy does not equal the ATT as holds true in the standard Difference-in-Differences model for continuous outcomes. While adding heterogeneity to standard Probit or Logit models has been done using a control function approach (Petrin and Train, 2010), to our knowledge the proper extension of Differences-in-Differences models to binary outcome variables allowing for continuous covariates has not been discussed so far.

Further, neither Probit, Logit or OLS approaches account fully for heterogeneity. Specifically, they do not allow for possible correlation of heterogeneous unobservables. For example, they do not allow for correlation between heterogeneous time invariant unobservables and heterogeneous treatment effects. Ignoring these effects could lead to biased estimates of the ATT, incorrect standard error estimation and, thus, flawed inference.

To close this gap in the literature, we propose a nonparametric binary choice Difference-in-Differences model with heterogeneous treatment effects. Our model accounts both for the non-linearity that emerges from the binary outcome setting and for treatment effect heterogeneity via the introduction of random coefficients. These random coefficients allow us to estimate the distribution of treatment effects on the treated which allows us to estimate the quantile treatment effects. As in the original Difference-in-Differences set-up, our model allows for correlation between treatment status, treatment effect and time-invariant unobservables.

The main idea of the paper is analogous to continuous outcome Difference-in-Differences. The data is split into a treatment group, affected by a particular treatment, and a control group, not affected by this treatment. We observe units in the treatment and the control group both before the treatment occurs and after the treatment occurs. It is important that we observe the same units before and after treatment. Furthermore, it is essential that pre-treatment observations are not causally affected by later occuring treatment due to anticipations effects.

In order to identify the ATT in a Difference-in-Differences model, one needs to be able to compare the actual post-treatment outcomes in the treatment group with the counterfactual post-treatment outcomes in the treatment group had treatment not occurred. The former are generally directly observable in the data. The latter are not and must thus be identified off the Difference-in-Differences model. How is it done? One has to make some type of common trend assumption for the treatment and control group. Namely, outcomes would develop similarly over time in the treatment group and the control group if no treatment occurred. This allows us to use the control group to isolate the time trend and thus to predict a post-treatment counterfactual outcome for the treatment group, absent treatment.

Several additional layers of complications arise in our case. First, binary outcome variables force us to work with latent outcome variables instead of directly observable outcome variables. Second, to identify the whole distributions of the random coefficients we need to introduce special regressors as the variation in the binary outcome variables is not sufficient. Third, we switch to a nonparametric set-up.

**Applications:** As binary outcome variables are very common in microeconomic modeling, applications of our proposed model are manifold. They range from the classical labor economic issues of female labor force participation, fertility decisions, and preventive health care decisions all the way to individual consumer demand estimation as a central question of the industrial organization literature.

In the labor economics literature, the body of existing applications using Difference-in-Differences methods in the context of a binary outcome variable is large. Two recent examples are Staubli (2011) and Campolieti and Riddell (2012). Both investigate the effect of a change in disability policy on employment and disability enrollment in Austria and Canada, respectively. Furthermore, papers by Schönberg and Ludsteck (2014) and Bargain et al. (2012) explore the effects of policy changes on female labor force participation. Schönberg and Ludsteck (2014) use the expansion in maternity leave coverage in Germany as treatment, whereas Bargain et al. (2012) make use of the introduction of divorce laws in Ireland as an exogenous shifter.

Prifti and Vuri (2013) and Dyer and Fairlie (2004) are recent examples of papers studying the fertility decision and out-of wedlock births, respectively. Prifti and Vuri (2013) look at the effects of a chance in employment protection legislation in Italy, whereas Dyer and Fairlie (2004) compare outcomes in states with family caps with outcomes in states without family caps. Lastly, Gruber and Poterba (1994) estimate the effects of tax incentives on the decision to buy health insurance.

To the best of our knowledge, difference-in-differences methods have so far not been used in individual consumer demand and willingness-to-pay estimation. Our model would be well-suited for investigating the effect of a public policy on the willingness-to-pay for a public good via contingent valuation studies as in Lewbel et al. (2011). Further, it fits our model better than standard DiD models because one would expect individual willingness-to-pay to be correlated with individual treatment effect. The fact that we are able to identify the joint distribution of actual and counterfactual latent outcomes, i.e. actual and counterfactual willingness-to-pay in this context, could be of political relevance.

Our application will examine the effect of a Sugar-Sweetened Beverages (SSB) tax implemented in Cook County, Illinois. We will do a difference-in-differences estimation comparing individuals in Cook County to individuals in neighboring counties. Under reasonable assumptions, our estimator will evaluate the effect of the SSB tax on consumers' likelihood to pay for soda. We will be able to estimate many aspects of the heterogeneous effects of the tax as well as the ATT.

**Related Literature:** Difference-in-Differences models have a long tradition in applied microeconomics. Theoretical literature on Difference-in-Differences methods is however fairly limited. Lechner (2011) provides an extensive overview of Difference-in-Differences models for continuous outcome variables. Some recent papers have extended the classical model in several directions. Athey and Imbens (2006) propose a scale invariant version of the Difference-in-Differences model. Another model by Bonhomme and Sauder (2011) accounts for treatment effect heterogeneity and is thus most closely related to our model. However, Bonhomme and Sauder (2011) along with all other papers cited above, do not allow for binary outcome variables.

Our paper also contributes to the literature on special regressors. Special regressors are exogenous regressors with full or large support that have been suggested to introduce additional observable variation into e.g. binary choice models thus helping identification of parameters or distributions of interest. Leading references are the papers by Lewbel (2000) and Dong and Lewbel (2015) and the survey article by Lewbel (2014).

Furthermore, our model features nonparametrically identified and estimated random coefficients and thus relates to the random coefficients literature. Hoderlein et al. (2010) discuss nonparametric identification of a linear random coefficient model, whereas Ichimura and Thompson (1998) and Gautier and Kitamura (2013) show nonparametric identification of a binary choice random coefficient model.

Lastly, there is a large strand of literature on consumer heterogeneity in discrete choice models used for demand estimation in industrial organization. Examples include parametric approaches as the random coefficients Logit model suggested in Berry et al. (1995b), as well as nonparametric approaches as discussed in Berry and Haile (2010) and Fox and Gandhi (2016).

**Outline of the Paper:** Section 2 focuses on the main identification result. We start with a discussion of the precise assumptions we require, and present and discuss the main result, which establishes the identification of the average treatment effect on the treated (ATT) in the binary choice Difference-in-Differences Model with heterogeneous treatment effects. Further, we present two extensions. Extension 1 shows identification of the joint distribution of the actual and counterfactual latent outcomes for the treatment group. Extension 2 outlines how covariates can be included in the basic model in two alternative ways. Section 3 proposes a sample counterpart estimator of our model. Section 4 documents the results from Monte Carlo simulations. Section 5 contains the results of an empirical application of our estimator on the SSB tax in Cook County, Illinois. Section 6 contains a summary and concluding remarks.

## 3.2 Identification

### 3.2.1 Basic Model without Covariates

**Notation and Assumptions:** Our proposed model makes use of the latent variable formulation and takes the following form for $t = 1, 2$.

$$\begin{aligned} Y_1^* &= B_1 - Z_1 \\ Y_2^* &= B_2 + B_3 D - Z_2 \\ B_2 &= B_1 + V_2 \\ Y_t &= \mathbf{1}\{Y_t^* < 0\} \end{aligned} \tag{3.2.1}$$

$Y_t$ denotes the binary (observed) outcome variable of interest. $Y_t^*$ denotes the latent outcome variable, which is unobservable. $\mathbf{1}\{\cdot\}$ represents the indicator function, such that $Y_t = 1$ if $Y_t^* < 0$, and $Y_t = 0$ otherwise. $D$ is a binary variable denoting whether an individual obtains treatment ($D = 1$) or belongs to the control group ($D = 0$). $B = (B_1, B_2, B_3)$ are random coefficients with unknown distribution satisfying the above restrictions. $V_2$ can be interpreted as a time trend plus potential shock at $t = 2$. $B_3$ denotes the effect of treatment on the latent outcome variable $Y_2^*$. $Z = (Z_1, Z_2)$ are special regressors.

**Example for Empirical Application:** Effect of an advertising campaign on the willingness to pay for a good. $D = 1$ denotes individuals who are exposed to the campaign (treatment group), $D = 0$ denotes individuals who are not exposed to the campaign (control group). Price, $Z_t$, is chosen randomly from a known distribution with large support in the course of the contingent valuation experiment conducted both before and after treatment occurs. At both time points individuals make decisions on whether they will purchase ($Y_t = 1$) the good or not ($Y_t = 0$) at price $Z_t$. $Y_t$ is observed by the econometrician. $B_1$

represents the individual preference heterogeneity, i.e. willingness to pay for the good and $B_3$ the treatment effect. $V_2$ summarizes both trends between time periods in demand for the good as well as second period shocks. The latent outcome variable $Y_t^*$ can thus be directly interpreted as the offered price minus the individual's willingness to pay, which would be their utility from purchasing that good. If the willingness to pay exceeds the price, the individual will purchase the good. [1].

We impose the following assumptions on the model:

**Assumption A.3.1** *Let* $(\Omega, F, P)$ *be a complete probability space on which are defined the random vectors* $(B_1, B_2, B_3, V_2) : \Omega \to \mathcal{B}_1 \times \mathcal{B}_2 \times \mathcal{B}_3 \times \mathcal{V}_2$, $\mathcal{B}_1 \subseteq \mathbb{R}$, $\mathcal{B}_2 \subseteq \mathbb{R}$, $\mathcal{B}_3 \subseteq \mathbb{R}$, $\mathcal{V}_2 \subseteq \mathbb{R}$, *and* $(D, Y_t, Y_t^*, Z_t) : \Omega \to \mathcal{D} \times \mathcal{Y} \times \mathcal{Y}^* \times \mathcal{Z}$, $\mathcal{D} = \{0,1\}$, $\mathcal{Y} = \{0,1\}^2$, $\mathcal{Y}^* \subseteq \mathbb{R}^2$, $\mathcal{Z} \subseteq \mathbb{R}^2$, $t = 1, 2$, *such that for* $t = 1, 2$, $(i)$

$$\begin{aligned} Y_1^* &= B_1 - Z_1 \\ Y_2^* &= B_2 + B_3 D - Z_2 \\ B_2 &= B_1 + V_2 \\ Y_t &= \mathbf{1}\{Y_t^* < 0\} \end{aligned}$$

*where* $(ii)$ *realizations of* $(D, Y_t, Z_t)$ *are observable, whereas those of* $(Y_t^*, B_1, B_2, B_3, V_2)$ *are not.*

**Assumption A.3.2** *There is no pre-treatment effect in period 1, i.e. there is no causal effect of* $D$ *on* $Y_1^*$.

**Assumption A.3.3** $V_2 \perp D, B_1, Z$.

**Assumption A.3.4** $Z \perp B_1, B_3 | D$.

[1] Please note that we are laxly calling $Z_t$ the price here, where $Z_t$ is actually the negative price. Similarly for the willingness to pay $B_1$. This is just to make the example more intuitive and comes at no cost, as we can always arbitrarily recode our binary $Y_t$ variable. Further, the willingness to pay and treatment effect variables ($B_1$ and $B_3$) are actual ratios of willingness to pay and treatment effect and the price elasticity, since the price variable has no coefficient.

**Assumption A.3.5** *$Z$ has full support and the support of $Z_2 - Z_1$ spans the support of $V_2$.*

**Discussion of Assumptions:** Assumption **A.3.1** formally specifies the data generating process discussed at the beginning of this section. The special regressors $Z$ are crucial for the identification of the distributions of the random coefficients in the model, as they introduce observable variation, where the outcome variables can only take on one of two values. The effect of $Z$ on the latent variable is assumed to be of known sign. The corresponding coefficients are normalized to one ensuring identification of the remaining coefficients of the model[2].

Assumption **A.3.2** is one of the classical Difference-in-Differences assumptions. If individuals in the treatment group somehow anticipate the treatment and react to this anticipation, our estimate of the average treatment effect on the treated (ATT) will be biased. What does this assumption say in the context of our example? If the news reports on the advertising campaign before the campaign begins, the treated individuals might exhibit higher willingness to pay already *before* treatment actually takes place. Thus there would be an anticipation effect, the assumption would be violated and our estimated ATT would be biased towards zero.

Next, Assumption **A.3.3** contains three independence assumption. First, $V_2 \perp D$ replaces another classical Difference-in-Differences assumption of common time trends in the treatment and control groups. The time trend $V_2$ is random in our model and needs to be independent of treatment status $D$. As in classical Difference-in-Differences we will identify the time trend from the control group and then use it to construct the counterfactual outcomes of the treatment group had they not been treated. If the time trends are different in the two groups, then the Difference-in-Differences identification strategy breaks

[2] Please see Appendix 3.7.2 for a detailed discussion of the restrictions we impose on the coefficients on $Z$.

down. What does this condition mean in terms of our purchase decision example? If the treatment group was also affected by a different sales campaign besides the advertising campaign we are examining, but the control group was not, then the time trend in the two groups absent our treatment of interest would likely not look the same. Our strategy of identifying the time trend of the treatment group off the control group would thus fail.

Second, $V_2 \perp B_1$ is a technical assumption that becomes necessary for our identification proof as both the intercept $B_1$ and the time trend $V_2$ are random in our model. If $V_2 \perp B_1|D$ does not hold, we cannot perform convolution with $f_{B_1|D=1}$ and $f_{V_2|D=1}$ to obtain $f_{B_1+V_2|D=1}$ in our identification proof. Let's look at this assumption in the light of our example. $B_1$ represents the individual's willingness to pay. $V_2$ denotes the time trend that includes preference changes along with personal preference shocks. Our assumption says that any preference changes from period 1 to period 2 are unrelated to individual's initial willingness to pay. This might be violated if individuals with a higher initial willingness to pay live clustered in certain areas. Thus, individuals living in each of these areas might change their preference to be more aligned with their neighbors between the pre-treatment and post-period time periods. Thus, preferences might change systematically differently for high versus low willingness to pay individuals.

The last independence $V_2 \perp Z$ imposes that the time trend has to be independent of the special regressors in either of the periods. In our example this means that the offered prices cannot depend on the time trend, i.e. it cannot be the case that individuals that experience a demand shock between the two time periods are offered higher or lower prices than other individuals. This assumption would easily hold if we could randomly assign prices to our individuals as in Lewbel et al. (2011). This exogeneity assumption may be difficult to satisfy in most other empirical applications, but Extension 3 discussed in section 3.2.5

allows for a control function approach that allows for an endogenous special regressor.

The next assumption, **A.3.4**, is another exogeneity assumption on the special regressors $Z$. The special regressors $Z$ need to be independent of the random coefficients $B_1, B_3$, conditional on the treatment status $D$. What does this assumption mean in terms of our example? Within treatment and control groups, the initial willingness to pay $B_1$, and the treatment effect $B_3$ are not correlated with prices $Z$. Again, this assumption is not trivially valid, but Extension 3 in Section 3.2.5 allows for more flexible empirical applications. This assumption – combined with the third independence assumption in **A.3.3** – is needed in our identification proof, as this allows us to scratch the conditioning on $Z$. If the $Z$'s were still conditioned on, we would not be able to integrate over the $Z$'s and thus our proof would break down.

Finally, the last assumptions, **A.3.5** are support assumptions on $Z$. In our willingness to pay example, these assumptions are met as long as prices vary sufficiently over time. This may not be the case in empirical applications as stores will keep the price near equilibrium, but Extension 4 discussed in Section 3.3.5 allows for a semi-parametric approach that allows for special regressors with limited support. The first part of assumption **A.3.5** is a classical special regressor assumption. To help overcome the fact that $Y$ exhibits only very little variation due to its binary character a variable $Z$ with plenty of variation is needed[3]. See Lewbel (2000), Lewbel (2014), or Dong and Lewbel (2015) for more information on how special regressors can be used to ensure identification in a binary choice model. The second part of assumption **A.3.5** is a necessary condition for our identification proof to go through, as the distribution of the time trend $V_2$ is broadly speaking identified as a difference of control group

[3] Please note that the full support assumption on $Z$ is a sufficient condition for the identification of the ATT under arbitrary distributions of the random coefficients. If some combinations of random coefficients have distributions with restricted support, this assumption can we weakened accordingly. Moreover, $Z$ could actually be a linear combination of more than one variable. See Extension 2 for more information on how to apply this idea in practice.

outcomes $Z$ between the two observed time periods before and after treatment.

**The Main Result:** The assumptions introduced above now allow us to identify the object of interest, the ATT. Our result is as follows:

**Theorem 3.1.** *Let Assumptions* ***A.3.1*** *–* ***A.3.5*** *hold. Then*

$$ATT \quad = \quad \int CATT(z_2) \quad F_{Z_2}(dz_2),$$

*where*

$$\begin{aligned} CATT(z_2) \quad &= \quad P^1[Y_2 = 1|D = 1, Z_2 = z_2] - P^0[Y_2 = 1|D = 1, Z_2 = z_2] \\ &= \quad F_{B_1+V_2+B_3|D}(z_2; 1) - F_{B_1+V_2|D}(z_2; 1) \\ &= \quad F_{B_1+V_2+B_3|D}(z_2; 1) - \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{B_1+V_2,B_1|D}(y + x, x; 0) dx F_{B_1|D}(z_2 - y; 1) dy, \end{aligned}$$

*and*

$$\begin{aligned} F_{B_1+V_2+B_3|D}(Z_2; 1) \quad &= \quad P[Y_2 = 1|D = 1, Z_2 = z_2], \\ f_{B_1+V_2,B_1|D=0} \quad &= \quad \partial_{Z_1} \partial_{Z_2} P[Y_2 = 1, Y_1 = 1|D = 0, Z = z], \\ F_{B_1|D=1} \quad &= \quad P[Y_1|D = 1, Z_1 = z_1]. \end{aligned}$$

**Remark: 2.1 - Discussion of Theorem 3.1:** Our main result establishes that the ATT is identified in our model. The ATT is usually the primary object of interest in the treatment effect literature, as it measures the average treatment effect on the actually treated population. In our example above, the

ATT measures the average effect of the advertising campaign on the probability that the households consume the product among treated households, i.e. households who were subject to the advertising campaign. The proof of Theorem 3.1 can be found in Appendix 3.7.1. Note that if $F_{B_1+V_2+B_3|D}(z_2;1)$ and $F_{B_1+V_2|D}(z_2;1)$ are uniform distributions, our model breaks down to a simple Linear Probability Model.

### 3.2.2 Extension 1: Joint Distribution of Latent Variables

Even though the focus of our paper lies in identifying the ATT in this model, as shown in the previous section, we can actually identify the joint distribution of all random coefficients in the model conditional on being treated. For this we need to go through characteristic functions.

**Theorem 3.2.** *Let Assumptions **A.3.1** – **A.3.5** hold. Then*

$$\begin{aligned}\Phi_{B_1,B_3,V_2|D=1}(\sigma,t,r) &= E\left\{\exp\left[i(\sigma B_1+tB_3+rV_2)\right]|D=1\right\}\\ &= E\left\{\exp\left[i(\sigma B_1+tB_3)\right]|D=1\right\}\times E\left\{\exp(irV_2)|D=1\right\},\end{aligned}$$

*where*

$$E\left\{\exp(irV_2)|D=1\right\} = E\left\{\left.\frac{\exp\left[ir(Z_2-Z_1)\right]\psi(Z_1,Z_2,0)}{f_{Z|D}(Z;0)}\right| D=0\right\},$$

*and*

$$E\left\{\exp\left[i(\sigma B_1+tB_3)\right]|D=1\right\} = \frac{E\left\{\left.\frac{\exp[i(\sigma Z_1+t(Z_2-Z_1))]\psi(Z_1,Z_2,1)}{f_{Z|D}(Z;1)}\right| D=1\right\}}{E\left\{\left.\frac{\exp[it(Z_2-Z_1)]\psi(Z_1,Z_2,0)}{f_{Z|D}(Z;0)}\right| D=0\right\}},$$

*with*

$$\psi(z_1,z_2,d) = \partial_{z_1}\partial_{z_2}P\left[Y_2=1,Y_1=1|D=d,Z=z\right].$$

**Remark: 2.2 - Discussion of Theorem 3.2:** This theorem establishes that the model proposed by us is fully identified. The proof of Theorem 3.2 can be found in Appendix 3.7.1. Starting from the above result, it is easy to show that the joint distribution of the actual and counterfactual latent variables in the post-treatment period are also identified for the treatment group, i.e. $f_{B_1+V_2+B_3,B_1+V_2|D=1}$ is identified[4]. In our example above, this distribution would capture the joint distribution of the actual and counterfactual willingness to pay of individuals in the treatment group, i.e. individuals who were subject to the anti-smog campaign. This could be of high interest to policy makers, e.g. when deciding on the pricing of public goods.

### 3.2.3 Extension 2A: Conditioning on Covariates

Our proposed model can be easily adjusted to include covariates as is common in applied research by conditioning everything on the covariates. This can even help to relax some of our assumptions of the previous section, as these will now only have to hold conditional on the covariates. Let us go back to our example to illustrate this point.

When e.g. demand trends are different in treatment and control groups, our assumption A.3.3 does not hold, as $V_2 \not\perp D$, and identification in our basic model without covariates breaks down. But what if demand trends are really only different for people on different income levels and the differential time trends in treatment and control group are due to the different income distributions in the two groups? Then we get that among people with the same income environmental trends are independent of treatment status, i.e. $V_2 \perp D|X$, and we can identify our model conditional on $X$.

Alternatively, when $B_1 \perp Z|D$ is not satisfied, as e.g. within treatment and control groups, individuals with higher willingness to pay are offered higher

[4] For example, obtain $f_{B_1,V_2,B_3|D=1}$ via multivariate inverse Fourier transform. Then, apply change of variables to obtain $f_{B_1+V_2+B_3,B_1+V_2,V_2+B_3|D=1}$. Lastly, get $f_{B_1+V_2+B_3,B_1+V_2|D=1}$ via integration.

prices, then again identification in our basic model without covariates breaks down. But what if the above dependency was fully driven by a third factor, say income? Then, once we condition on income, we will not see any relationship between willingness to pay and prices among individuals of the same income in treatment and control groups, respectively, i.e. $B_1 \perp Z|D, X$, and again we can identify our model conditional on $X$.

**Model:**

$$\begin{aligned} Y_1^* &= B_1(X,\omega) - Z_1 \\ Y_2^* &= B_2(X,\omega) + B_3(X,\omega)D - Z_2 \\ B_2(X,\omega) &= B_1(X,\omega) + V_2(X,\omega) \\ Y_t &= \mathbf{1}\{Y_t^* < 0\} \end{aligned}$$

where $X$ denotes the random vector of observable covariates and $\omega$ denotes the unobservable random scalar.

**Assumption A.3.2.1** *There is no pre-treatment effect in period 1, i.e. there is no causal effect of $D$ on $Y_1^*$ conditional on $X$.*

**Assumption A.3.3.1** $V_2 \perp D, B_1, Z|X.$

**Assumption A.3.4.1** $B_1, B_3 \perp Z|D, X.$

**Assumption A.3.5.1** *$Z$ has full support and the support of $Z_2 - Z_1$ spans the support of $V_2$ conditional on $X$.*

### 3.2.4 Extension 2B: Exogeneous Covariates with Fixed Coefficients

Of course conditioning everything on the covariates also means that we obtain all distributions of the random coefficients conditional on the covariates, i.e. we receive a different set of distributions for every possible set of values $x$. If the set of possible values that the covariate vector $X$ can take on becomes large, e.g. due to a continuous variable in $X$, conditioning on all covariates might become infeasible in practice. In this case we suggest to include exogenous covariates with fixed coefficients into the model.

**Model:**

$$\begin{aligned} Y_1^* &= B_1 + X_1\gamma_1 - Z_1 \\ Y_2^* &= B_2 + B_3 D + X_2\gamma_2 - Z_2 \\ B_2 &= B_1 + V_2 \\ Y_t &= \mathbf{1}\{Y_t^* < 0\} \end{aligned}$$

where $X_1$ and $X_2$ denote the respective sets of obervable covariates, and $\gamma_1$ and $\gamma_2$ denote the vectors of fixed coefficients.

### 3.2.5 Extension 3: Using a Control Function with an Endogenous Special Regressor

Finding an exogenous special regressor can be difficult in many scenarios, so it may be necessary to use an instrumental variable. For example, in demand estimation, price seems like a logical special regressor but has some endogeneity concerns that are typically solved by using instrumental variables. Controlling for this type of endogeneity opens the door to many more empirical applications.

We will use a control function approach first introduced by Heckman and Robb (1985). This approach is helpful in our case because it is powerful when dealing with heterogeneous effects and is useful for nonparametric models Arellano and Bond (1991). Thus, we assume the following model:

**Model:**

$$\begin{aligned}
Y_1^* &= B_1 - Z_1 \\
Y_2^* &= B_2 + B_3 D - Z_2 \\
B_2 &= B_1 + V_2 \\
Y_t &= \mathbf{1}\{Y_t^* < 0\} \\
Z_t &= G(W_t) + \epsilon_t
\end{aligned}$$

Where $W_t$ is the instrumental variable and $G(W_t)$ is the control function. Thus, we can adjust our estimator as long as $W$ holds standard IV assumptions like $E[\epsilon|W] = 0$. Thus, we now only need the following assumptions to hold rather than assuming that $Z$ is exogenous:

**Assumption A.3.3.2** $V_2 \perp D, B_1, W.$

**Assumption A.3.4.2** $B_1, B_3 \perp W|D.$

**Assumption A.3.5.1** *$W$ has full support and the support of $W_2 - W_1$ spans the support of $V_2$ conditional on $\epsilon$.*

In our previous example, this would allow us to use different instrumental variables for price, such as Hausman Instruments (Hausman et al., 1994), the average prices in neighboring cities of the good. These Hausman prices must be independent of the consumer's willingness to pay for a good in the pre-treatment or post-treatment time period as well as the time trend of demand, which is similar to standard assumptions made in the demand estimation literature where these are commonly used.

## 3.3 Estimation

### 3.3.1 Basic Model without Covariates

Estimation is performed via sample counterparts.

$$\widehat{ATT} = \frac{1}{n}\sum_{i=1}^{n} \widehat{CATT}(z_{2i}),$$

where

$$\widehat{CATT}(z_2) = \hat{F}_{B_1+V_2+B_3|D}(z_2;1) - \hat{F}_{B_1+V_2|D}(z_2;1).$$

$\hat{F}_{B_1+V_2+B_3|D=1}$ can be directly estimated from the data via nonparametric regression:

$$\hat{F}_{B_1+V_2+B_3|D}(z_2;1) = \hat{P}[Y_2 = 1|D = 1, Z_2 = z_2].$$

Estimating the counterfactual CDF $\hat{F}_{B_1+V_2|D=1}$ involves generating data and then estimating the generated data's empirical CDF.

Consider:

$$\begin{aligned} F_{B_1+V_2|D}(z_2;1) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbf{1}\{b_1 + v_2 < z_2\} f_{B_1,V_2|D}(b_1, v_2; 1) db_1 dv_2 \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbf{1}\{b_1 + v_2 < z_2\} f_{B_1|D}(b_1;1) f_{V_2|D}(v_2;1) db_1 dv_2 \end{aligned}$$

And thus for generated data $b_{1i}$ and $v_{2i}$:

$$\hat{F}_{B_1+V_2|D}(z_2;1) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{b_{1i} + v_{2i} < z_2\}$$

How to generate $b_{1i}$ and $v_{2i}$?

1. Generating $b_{1i}$:

   (a) Estimate $F_{B_1|D=1}$ via local constant regression of $P[Y_1 = 1|D = 1, Z_1 = z_1]$.

(b) Perform inverse transform sampling to obtain generated data points $b_{1i}$.

2. Generating $v_{2i}$[5]:

   (a) Estimate $\partial_{z_1} F_{B_1+V_2,B_1|D=0}(z_2, z_1)$ via local linear regression of $P[Y_2 = 1, Y_1 = 1|D = 0, Z = z]$ (partial derivative).

   (b) Estimate $f_{B_1|D=0}(z_1)$ via local linear regression of $P[Y_1 = 1|D = 0, Z_1 = z_1]$ (first derivative).

   (c) Obtain $F_{V_2}(z_2-z_1) = F_{V_2|D=0}(z_2-z_1)$ as the ratio of $\partial_{z_1} F_{B_1+V_2,B_1|D=0}(z_2, z_1)$ over $f_{B_1|D=0}(z_1)$. Average over all $F_{V_2}(z_2 - z_1)$ with $z_2 - z_1 = v_2$ to obtain $F_{V_2}(v_2)$.

   (d) Perform inverse transform sampling to obtain generated data points $v_{2i}$.

### 3.3.2 Extension 2A: Conditioning on Covariates

Include $X$'s into all local polynomial regressions mentioned above and get all densities as before but conditional on $X$.

### 3.3.3 Extension 2B: Exogeneous Covariates with Fixed Coefficients

1. Estimate $\gamma_1$ and $\gamma_2$ semi-parametrically via average derivative estimator, e.g. Ichimura (1991), Klein & Spady (1993), etc.

2. Generate new $Z$ variables and proceed with estimation as before.

### 3.3.4 Extension 3: Using a Control Function with an Endogenous Special Regressor

1. Estimate $\widehat{G}(W)$ non-parametrically from the equation $Z = \widehat{G}(W) + \epsilon$.

[5] The intuition behind this estimation is that the variation in $Z$ moves the same as the variation in $V_2$. Thus, the different movements of $Z$ can allow us to trace what happened with $V_2$ across the distribution

2. Use $\widehat{G}(W)$ and $Z$ to estimate $\widehat{\epsilon}$

3. Include $\widehat{\epsilon}$'s into all local polynomial regressions mentioned above and get all densities as before but conditional on $\widehat{\epsilon}$.

### 3.3.5 Extension 4: Semi-parametric Estimation for Special Regressors with Limited Support

In applications it is often not plausible to find special regressors $Z$ with full support. In this section, we thus extend our model to handle special regressors with compact support via a semi-parametric approach. The basic argument is an extrapolation argument. First, we estimate non-parametrically the two CDFs of interest $\hat{F}_{B_1|D=1}$ and $\hat{F}_{V_2}$ as described above, but only on a compact support. In a second step, we then construct a minimum-distance parametric estimator on the limited support based on a known parametric distribution (with full support), e.g. the normal distribution. We can then use the estimates for the parameters of these known distributions to obtain estimates for our CDFs of interest outside the compact support provided by the special regressors $Z$. Once we put the non-parametric and parametric part together, we can perform inverse transform sampling as if the special regressors had full support.

Formally, let $Z_1 \in [\underline{z}_1, \overline{z}_1]$ and $Z_2 \in [\underline{z}_2, \overline{z}_2]$. The process of generating $b_{1i}$ now needs an additional step between 1(a) and 1(b), as $\hat{F}_{B_1|D=1}$ is only obtained for $\underline{z}_1 < b_1 < \overline{z}_1$. To obtain $\hat{F}_{B_1|D=1}$ for $b_1 < \underline{z}_1$ and $b_1 > \overline{z}_1$, minimize the following criterion:

$$\min_{\theta} \sum_{i=1}^{N_1} \left[ \hat{F}_{B_1|D=1}(z_{1i}) - CDF(z_{1i}, \theta) \right]^2 ,$$

where $z_{1i}$ are the sample realizations of $Z_1$ for $D = 1$. For a normal distribution, this yields:

$$\min_{\mu,\sigma} \sum_{i=1}^{N_1} \left[ \hat{F}_{B_1|D=1}(z_{1i}) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(z_{1i}-\mu)/\sigma} \exp(-t^2/2)dt \right]^2 .$$

The process of generating $v_{2i}$ also needs an additional step between 2(c) and 2(d), as $\hat{F}_{V_2}$ is only obtained for $\underline{z}_2 - \overline{z}_1 < v_2 < \overline{z}_2 - \underline{z}_1$. To obtain $\hat{F}_{V_2}$ for $v_2 < \underline{z}_2 - \overline{z}_1$ and $v_2 > \overline{z}_2 - \underline{z}_1$, minimize the following criterion:

$$\min_{\theta} \sum_{i=1}^{N_0} \left[ \hat{F}_{V_2}(v_{2i}) - CDF(v_{2i}, \theta) \right]^2 ,$$

where $v_{2i}$ are the sample realizations of $Z_2 - Z_1$ for $D = 0$. What follows is analogous to generating $b_{1i}$.

This approach is less prone to misspecification than a fully parametric approach, as we only use a parametric distribution where we cannot obtain estimates non-parametrically in the estimation of the counterfactual CDF $\hat{F}_{B_1+V_2|D=1}$ due to limited support of the special regressors $Z$. The actual CDF $\hat{F}_{B_1+V_2+B_3|D=1}$ can always be estimated fully nonparametrically.

Note that you can use both Extension 3 and Extension 4 in our model. However, in order to properly, the CDF functions must be constructed conditional on $\epsilon_i$. We will demonstrate this in our empirical application.

### 3.3.6 Extension 5: Parametric Estimation for Special Regressors with Limited Support

An alternative approach to work around a limited support of the special regressors is to fully parametrize the estimation of our model. This means that we either have to assume parametric distributions directly for $B_1 + V_2 + B_3|D = 1$, $B_1|D = 1$, and $B_1 + V_2, B_1|D = 0$ or indirectly through distributional assumptions on the random coefficients including their dependence structure.

The easiest way for now seems to be to assume normal distributions for the three expressions above, two univariate normal distributions and one bivariate normal distribution. This can be extended later on. Once we know the three above mentioned quantities have normal distributions, we can find the respective parameters of these distributions via univariate and bivariate Probit.
For $B_1|D = 1$, e.g., we can write:

$$\begin{aligned} P[Y_1 = 1|D = 1, Z_1 = z_1] &= P[B_1 < z_1|D = 1, Z_1 = z_1] \\ &= P[B_1 < z_1|D = 1] \\ &= \Phi\left(\frac{z_1 - \mu}{\sigma}\right) \\ &= \Phi\left(-\frac{\mu}{\sigma} + \frac{1}{\sigma}z_1\right) \\ &= \Phi(\beta_0 + \beta_1 z_1) \end{aligned}$$

Once we have $\hat{F}_{B_1+V_2+B_3|D}(z_2; 1)$, $\hat{f}_{B_1+V_2,B_1|D}(z_2, z_1; 0)$, and $\hat{F}_{B_1|D}(z_1; 1)$, we can plug in the sample values for $z_2$ and integrate numerically to obtain $\widehat{CATT}(z_2)$ for all $z_2$.

$$\widehat{CATT}(z_2) = \hat{F}_{B_1+V_2+B_3|D}(z_2; 1) - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \hat{f}_{B_1+V_2,B_1|D}(y + x, x; 0)dx\hat{F}_{B_1|D}(z_2 - y; 1)dy$$

Take the sample average over all $z_2$ to obtain an estimate of the ATT.

Another advantange of the fully parametric approach is that covariates $X$ can be easily included.

How is this different from the Probit model that practitioners like to estimate for binary choice difference-in-differences? In terms of estimation, practitioners usually estimate only one Probit model of the following specification:

$$\hat{P}[Y = 1|D = d, T = t, X = x] = \Phi(\beta_0 + \beta_1 D + \beta_2 T + \beta_3 DT + \delta X),$$

where $Y$ and $X$ are the stacked observations of the pre-treatment and post-treatment periods, the dummy variable $D$ denotes treatment status and the dummy variable $T$ denotes whether the observation is pre-treatment or post-treatment. Following from the discussion by Puhani (2012), the proper ATT estimate is then:

$$\widehat{ATT} = \frac{1}{n}\sum_{i=1}^{n}\left(\Phi(\beta_0 + \beta_1 + \beta_2 + \beta_3 + \delta x_i) - \Phi(\beta_0 + \beta_1 + \beta_2 + \delta x_i)\right),$$

where $n$ is the number of observations in the treatment group post-treatment in the sample. Clearly, this quantity is different from the quantity above. What is the intuition? The standard Probit model does not allow for correlation between treatment status and time invariant unobservables, but our model does. This is why our model requires some extra steps. To obtain the counterfactual distribution for the treatment group had the treatment not occurred, one has to use information on the time trend $V_2$ which has to be identified off the control group first.

## 3.4 Monte Carlo Simulations

We run Monte Carlo simulations of our model with different set-ups. We start off with a low dependence DGP, with $P(Y_t = 1)$ around one half in expectation. Then, we run different variations increasing $P(Y_t = 1)$ (Variation a), increasing $P(Y_t = 1)$ further (Variation b), and decreasing $P(Y_t = 1)$ (Variation c) by changing the mean of the distributions of the random coefficients $B_1$ and $B_3$. Next, we manipulate the binomial distribution of the treatment indicator D to both a lower probability (Variation d) and a higher probability (Variation e).

A second set of simulations assumes a higher dependence structure in the DGP. In particular, the random coefficients $B_1$ and $B_3$, as well as the time trend plus potential shock in the second period $V_2$ now follow different distributions for treatment and control group. In a first variation of this scenario, we extend this high dependence to the special regressors $Z$ as well and make their distribution different for treatment and control group (Variation a). In a second variation, we introduce dependence between the random coefficient $B_3$, measuring the heterogeneous treatment effect on the latent outcome variable, and the time trend plus potential second period shock $V_2$ (Variation b).

We run simulations with 1000 repetitions and sample sizes $N = 1,000$. We estimate the ATT the following ways: we use our standard non-parametric estimator (CHW), our estimator using the semi-parametric specification as outlined in Extension 4 (CHW-P), a basic linear probability model (LPM), a standard Probit model, and a Oracle model. The oracle model is estimated by taking the distance between the true and counterfactual distributions of the latent variable based on the unobserved random coefficients. Bandwidths for the non-parametric estimations are determined via grid search. Let us recall the model:

$$\begin{aligned} Y_1^* &= B_1 - Z_1 \\ Y_2^* &= B_2 + B_3 D - Z_2 \\ B_2 &= B_1 + V_2 \\ Y_t &= \mathbf{1}\{Y_t^* < 0\} \end{aligned}$$

### 3.4.1 Simulation Set-Up 1 (Low Dependence DGP)

**Distribution of unobservables:**

$$(B_1, B_3, V_2) \sim N(\mu_0, \Sigma_0) \quad \text{with} \quad \mu_0 = \begin{pmatrix} -0.5 \\ -0.5 \\ 1 \end{pmatrix} \quad \text{and} \quad \Sigma_0 = \begin{pmatrix} 1 & 0.3 & 0 \\ 0.3 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

**Distribution of observables:**

$$(Z_1, Z_2) \sim N(\mu_Z, \Sigma_Z) \quad \text{with} \quad \mu_Z = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \text{and} \qquad \Sigma_Z = \begin{pmatrix} 4 & 0.5 \\ 0.5 & 4 \end{pmatrix}$$

$D$ is binomial with probability 0.5.

**Variations:**

(a) $\mu_0$ replaced with $\mu_0^a = \begin{pmatrix} 1.5 \\ 1.5 \\ 1 \end{pmatrix}$

(b) $\mu_0$ replaced with $\mu_0^b = \begin{pmatrix} 2.5 \\ 2.5 \\ 1 \end{pmatrix}$

(c) $\mu_0$ replaced with $\mu_0^c = \begin{pmatrix} -2.5 \\ -2.5 \\ 1 \end{pmatrix}$

(d) $D$ is binomial with probability 0.3.

(e) $D$ is binomial with probability 0.8.

### 3.4.2 Simulation Set-Up 2 (Higher Dependence DGP)

**Distribution of unobservables:** $(B_1, B_3, V_2) \sim N(\mu^d, \Sigma^d)$

$$\text{with} \quad \mu^1 = \begin{pmatrix} -1.5 \\ -1.5 \\ 1 \end{pmatrix} \quad \text{and} \quad \Sigma^1 = \begin{pmatrix} 1 & 0.3 & 0 \\ 0.3 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{for} \quad D = 1,$$

$$\text{and} \quad \mu^0 = \begin{pmatrix} -0.5 \\ -0.5 \\ 1 \end{pmatrix} \quad \text{and} \quad \Sigma^0 = \begin{pmatrix} 1 & 0.3 & 0 \\ 0.3 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{for} \quad D = 0.$$

**Distribution of observables:**

$$(Z_1, Z_2) \sim N(\mu_Z, \Sigma_Z) \quad \text{with} \quad \mu_Z = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \text{and} \qquad \Sigma_Z = \begin{pmatrix} 4 & 0.5 \\ 0.5 & 4 \end{pmatrix}$$

$D$ is binomial with probability 0.5

**Variations:**

(a) Distribution of observables now different for $D = 0$ and $D = 1$: $(Z_1, Z_2) \sim N(\mu_Z^d, \Sigma_Z^d)$

$$\text{with} \quad \mu_Z^1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \text{and} \qquad \Sigma_Z^1 = \begin{pmatrix} 4 & 0.5 \\ 0.5 & 4 \end{pmatrix} \qquad \text{for} \qquad D = 1$$

$$\text{and} \quad \mu_Z^0 = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix} \qquad \text{and} \qquad \Sigma_Z^0 = \begin{pmatrix} 6 & 0.5 \\ 0.5 & 6 \end{pmatrix} \qquad \text{for} \qquad D = 0$$

(b) Distribution of unobservables with higher dependence structure:

$$(B_1, B_3, V_2) \sim N(\mu^d, \Sigma^d)$$

$$\text{with} \quad \mu^1 = \begin{pmatrix} -0.3 \\ -0.7 \\ 1 \end{pmatrix} \quad \text{and} \quad \Sigma^1 = \begin{pmatrix} 1 & 0.3 & 0 \\ 0.3 & 1 & 0.1 \\ 0 & 0.1 & 1 \end{pmatrix} \quad \text{for} \quad D = 1,$$

$$\text{and} \quad \mu^0 = \begin{pmatrix} -0.5 \\ -0.5 \\ 1 \end{pmatrix} \quad \text{and} \quad \Sigma^0 = \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0.1 \\ 0 & 0.1 & 1 \end{pmatrix} \quad \text{for} \quad D = 0.$$

### 3.4.3 Simulation Results

Below are the results of simulations for sample size $N = 1,000$. Our model does consistently better than LPM and Probit, sometimes even better than an oracle estimator that estimates the distributions directly off the (in practice unobservable) realizations of random coefficients via empirical CDF. Gains of the CHW estimator compared to LPM and Probit seems to be mainly due to smaller biases. The gains for the CHW-P estimator appear to be smaller than the CHW estimator for most simulations but there are still significant improvements to the bias of the estimator.

Note that the Oracle estimates are biased in the same direction as the LPM and Probit estimates. This is because it generally suffers from the same bias as those estimates: they do not allow for correlation between treatment effect and time invariant unobservables. These correlations seem realistic in our applications to consumer demand and are thus included in each of our simulations.

We further show examples of simulations of the CATT random coefficients compared to the oracle and true CATT values in Figure 3.1. Note that these are singular simulations so they have higher variance than the average of our bootstrap estimates. The CHW and CHW-P estimates are generally very close and generally have similar shape and density as the true estimates.

Table 3.1: Simulation Results

| Set-up | Category | CHW | CHW-P | LPM | Probit | Oracle |
|---|---|---|---|---|---|---|
| 1 | MSE | 0.0013 | 0.0013 | 0.0023 | 0.0021 | 0.0013 |
| 1 | Bias | 0.0175 | -0.0090 | 0.0368 | 0.0361 | 0.0359 |
| 1 | Var | 0.0010 | 0.0012 | 0.0010 | 0.0008 | 0.0001 |
| 1a | MSE | 0.0005 | 0.0007 | 0.0012 | 0.0008 | 0.0005 |
| 1a | Bias | 0.0010 | -0.0111 | -0.0218 | -0.0190 | -0.0207 |
| 1a | Var | 0.0005 | 0.0006 | 0.0007 | 0.0004 | 0.0000 |
| 1b | MSE | 0.0011 | 0.0008 | 0.0010 | 0.0009 | 0.0006 |
| 1b | Bias | 0.0165 | 0.0108 | 0.0235 | 0.0234 | 0.0230 |
| 1b | Var | 0.0009 | 0.0007 | 0.0006 | 0.0003 | 0.0000 |
| 1c | MSE | 0.0015 | 0.0016 | 0.0040 | 0.0038 | 0.0032 |
| 1c | Bias | -0.0229 | -0.0276 | 0.0580 | 0.0547 | 0.0559 |
| 1c | Var | 0.0010 | 0.0023 | 0.0006 | 0.0008 | 0.0001 |
| 1d | MSE | 0.0013 | 0.0020 | 0.0023 | 0.0021 | 0.0014 |
| 1d | Bias | 0.0160 | 0.0186 | 0.0347 | 0.0341 | 0.0359 |
| 1d | Var | 0.0011 | 0.0016 | 0.0011 | 0.0009 | 0.0001 |
| 1e | MSE | 0.0017 | 0.0024 | 0.0027 | 0.0026 | 0.0013 |
| 1e | Bias | 0.0190 | -0.0164 | 0.0375 | 0.0378 | 0.0360 |
| 1e | Var | 0.0013 | 0.0021 | 0.0013 | 0.0012 | 0.0000 |
| 2 | MSE | 0.0025 | 0.0035 | 0.0084 | 0.0048 | 0.0072 |
| 2 | Bias | 0.0218 | 0.0262 | 0.0869 | 0.0633 | 0.0845 |
| 2 | Var | 0.0021 | 0.0028 | 0.0008 | 0.0008 | 0.0001 |
| 2a | MSE | 0.0025 | 0.0035 | 0.0050 | 0.0054 | 0.0071 |
| 2a | Bias | -0.0038 | -0.0024 | 0.0648 | 0.0672 | 0.0838 |
| 2a | Var | 0.0024 | 0.0035 | 0.0008 | 0.0009 | 0.0001 |
| 2b | MSE | 0.0014 | 0.0016 | 0.0037 | 0.0036 | 0.0026 |
| 2b | Bias | 0.0215 | -0.0039 | 0.0538 | 0.0532 | 0.0499 |
| 2b | Var | 0.0009 | 0.0015 | 0.0008 | 0.0008 | 0.0001 |

This table compares simulation results for several estimators: CHW, CHW-P, LPN, and probit. The sample size is $N = 1000$.

## 3.5 Empirical Application

In this section, we discuss all matters pertaining to our empirical implementation where we examine whether the taxes on Sugar Sweetened Beverages (SSBs)

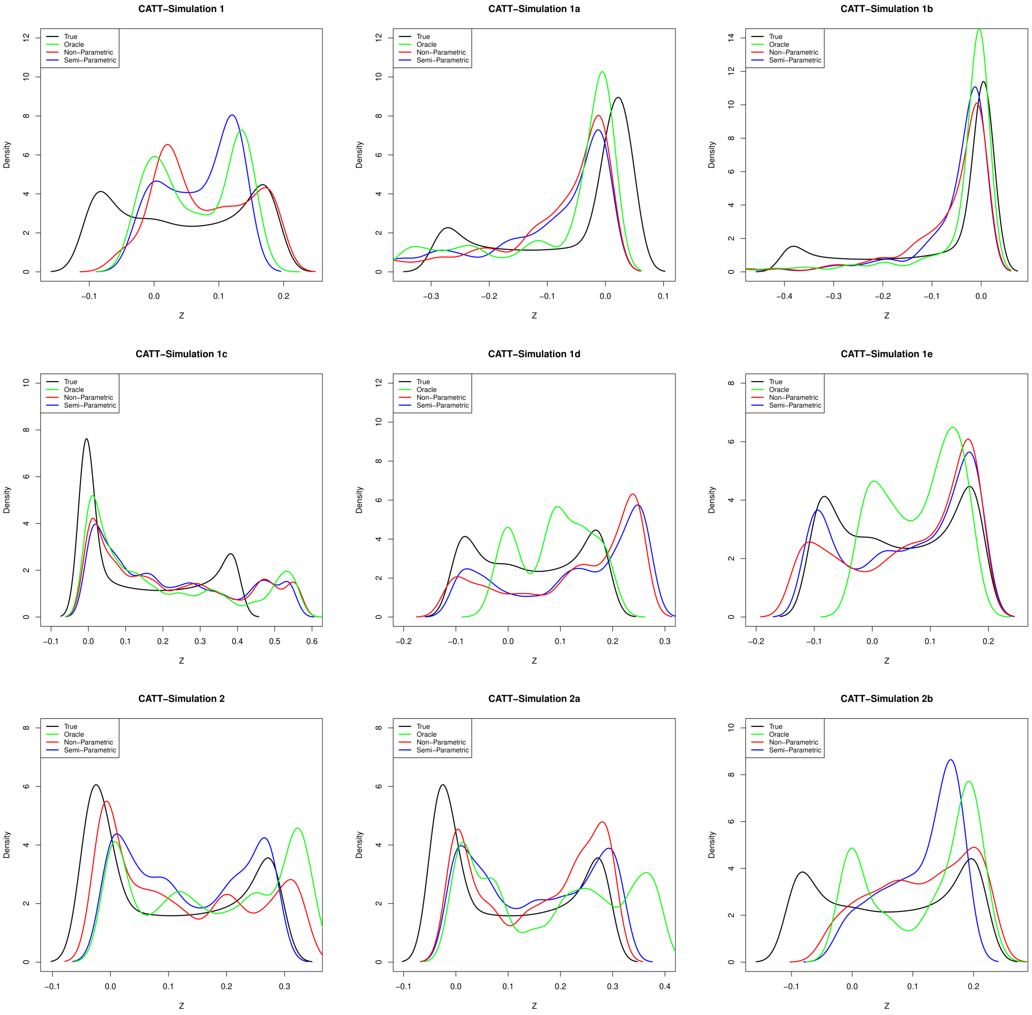

Figure 3.1: These graphs show the distribution of CATT for each of our simulations. We graph the true CATT values, oracle CATT values, CHW CATT values and CHW-P CATT values.

effected the fraction of individuals that consumed soda in Cook County, Illinois. This is an evaluation of the extensive effects of the tax: the impact of the tax on whether or not consumers purchased soda. We give an overview of the tax implementation, examine the data, discuss the current literature on the issue, and present the results.

### 3.5.1 SSB Tax

SSB tax was first implemented in Berkeley, California in November 2014 which specifically a one-cent-per-ounce soda . Starting in 2017, it was followed by many other cities implementing soda tax of different levels, including Philadelphia, PA, Oakland, CA, Albany, CA, Boulder, CO, San Francisco, CA and Seattle, WA. Since the tax level was different for most of these cities, we will examine the effect of the one-cent-per-ounce SSB tax implemented in Cook County, Illinois where the majority of the population lives in Chicago and thus is the largest single implementation of a SSB tax in the US.

The tax was passed into law in November 2016, and was expected to go into effect in July 1, 2017. However, on June 27th, the Illinois Retail Merchants Association filed a suit to challenge the constitutionality of the tax. On July 28th, the lawsuit was dismissed and the tax was implemented at the beginning of August. This sudden implementation of the tax prevented most consumers from purchasing large amounts of soda right before the treatment started. The government decided to repeal the tax in October of 2017 and the tax expired on December 1, 2017.

For most customers, the tax was shown as a line-item on their receipt. Some retailers simply added the tax to the display price at checkout. This could decrease the impact of the tax on soda consumption Chetty et al. (2009). However, many large retail stores in the area added a disclaimer to soda beverages stating that the tax would be added at the register. While this setting could lead to a significant impact on consumer information on the tax, this tax was newsworthy and I will assume that all of our households knew about the tax when making their purchase decision. In the end, a full Difference-in-Difference analysis is necessary because people respond to this tax differently than a standard price change.

### 3.5.2 Data

We will look at the Nielsen Scanner Data which is available through the Kilts Center at the University of Chicago Booth School of Business[6]. We will focus our study on the year 2014 where there are about 2,000 households in our area of interest. This is a helpful dataset for estimating consumption behavior since it contains detailed information based on price and quantity of all retail purchases as well as detailed household characteristics for all of the consumers. The data contains a representative sample of households in the United States that use in-home scanners to record all of their purchases intended for personal, in-home use. Nielsen matches the product scanned by the household to the actual price of the store where the product was bought. Nielsen estimates that about 30% of household consumption is accounted for by these purchases.

I will be doing a difference-in-difference analysis comparing Cook County to other counties in the designated market area surrounding Chicago to determine how the SSB tax effected whether households purchased soda. Because the law suits happened to soon before the tax was implemented, and Figure 3.2 shown below, we do not find evidence of a significant increase in the number of individuals who purchased soda right before the soda tax was implemented (as well as the fact that the tax was authorized less than a week before it was

[6] Researcher(s) own analyses calculated (or derived) based in part on data from The Nielsen Company (US), LLC and marketing databases provided through the Nielsen Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the Nielsen data are those of the researcher(s) and do not reflect the views of Nielsen. Nielsen is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

implemented), so we will use July 2017 as our pre-period. Although there has been evidence that consumers purchase soda when it is cheap in bulk and store it until it is on sale again (Hendel and Nevo, 2006), we avoid this problem by focusing on monthy sales and whether or not households consumed soda.

Note that it is possible that the SSB tax in Cook County decreased the probability a household consumed soda in the suburbs, but there is no evidence of this in the data and it is unlikely since the purchases being measured are made at retail stores which likely occur at stores near where the consumers' homes. Using similar cities to Chicago is a possible alternative but that might lead to other demand shocks that would bias our results.

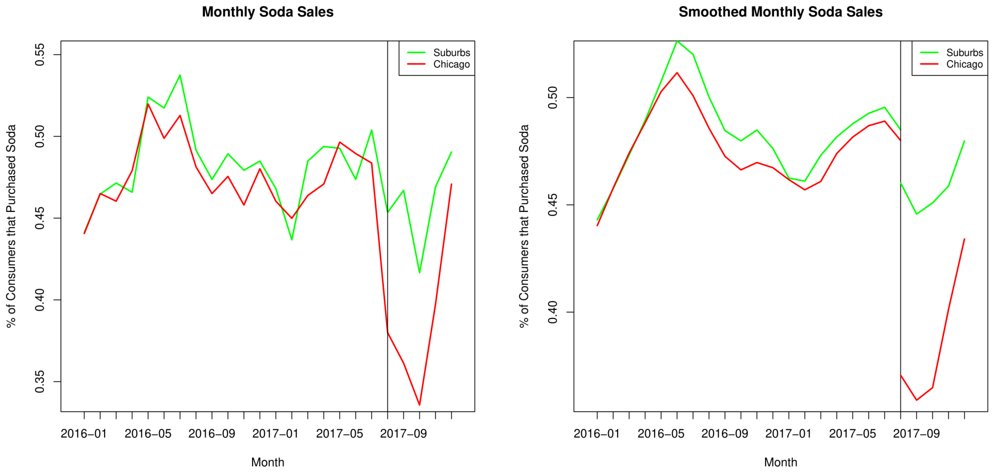

Figure 3.2: These graphs compare the fraction of consumers who purchased soda each month. The black vertical line is the implementation of the SSB tax in Cook County.

We aggregated the data to a monthly level such that period 1 is July 2017 and period 2 is September 2017. We chose September 2017 as the post-period to allow for the use of August 2017 as our instrumental variable to control for endogoneity, which we will discuss more later. We will use a measure of price deviation as our special regressor. Price is estimated using an aggregate price index called Stone-Lewbel (SL) cross section prices (see Lewbel (1989) and Hoderlein and Mihaleva (2008)). Generally, SL prices use the fact that within a category of goods (soda in our case), people have different tastes for the individual goods. Using standard aggregate price indices implicitly assumes that all individuals have identical Cobb Douglas preferences for all goods within a category, but SL prices allow all individuals to have heterogeneous Cobb Douglas preferences. This implies that the typical approach of using aggregate price indices is a restrictive case of using SL prices. For this reason, SL prices should always be used when possible.

There are a few concerns with the data. The data relies on participants successfully recording their purchases in their home, so they may suffer from some recording error. The specific issue that we should be concerned with is that consumers may consume a good when it is purchased so will not record the purchase when they return home. Einav et al. (2010) finds that consumable goods like soft drinks are likely to be consumed before getting home so are more likely to not be scanned. However, these errors only have minor effects on estimates. When compared to data from grocery store recorded sales, the data in Nielsen Homescan data matched 94% of the time Einav et al. (2010).

The major source of measurement error that is more concerning can come from the price rather than the quantity. Individuals record their purchases by scanning the items they buy when they get home. The individuals input the quantity they purchase and Nielsen matches it with the average price of the good at the store where they purchased it that week. This can lead to two types of errors. The first comes from the price changing in the middle of the week. These types of errors are approximately normally distributed. The second type of error comes from not including discounts from loyalty cards. Einav et al. (2010) examines a retailer used in the Homescan data which has loyalty cards and finds that loyalty cards are used in about 75-80% of the transactions. Further, this would bias our prices upwards, which when comparing Homescan data with data from the retailer finds that the prices used in the Homescan data is about 7% higher and the overall expenditure is 10% higher. On the other hand, these price measurement errors may be overestimated since some retailers do not have loyalty cards at all.

However, Homescan data errors are comparable to errors found in other commonly used data sets. Aguiar and Hurst (2007) finds that life-cycle pattern of household expenditures recorded in Homescan Data is consistent with those reported for food expenditures at home in Panel Study of Income Dynamics (PSID). Einav et al. (2010) finds that these issues are not more serious than those in any other consumption surveys like the Current Population Survey (CPS). Lin (2018) compares the fraction of expenditures on different categories of products in the Nielsen Homescan Data and finds the results consistent to results from the Consumer Expenditure Survey (CES).

### 3.5.3 Literature Review

Obesity is one of the most important health problems in the United States as well as other countries. Most soft drinks are high on sugar and excess sugar consumption is strongly linked with many diet-related diseases such as diabetes, cancers and heart disease World Health Organization (2015). Obesity leads to several hundred billion dollars spent on medical costs in the US annually, about 10-27 percent of all medical costs as shown in Finkelstein et al. (2009); Cawley et al. (2015). This is particularly relevant for policy makers since 88% of obesity related medical expenses as shown in Cawley and Meyerhoefer (2012). Thus, consumption of unhealthy food, such as soft drinks, can have a major economic impact. About 40% of sugar and 7% of total calories consumed by Americans come from soft drinks (United States Department of Agriculture, 2020; Allcott et al., 2019).

There is a strong interest in soft drink consumption among different groups of individuals (Dubois et al., 2019), such as children (Han and Powell, 2013) and low-income households (Drewnowski and Specter, 2004; Currie, 2009). Alcott et al. (2017) showed that even when controlling for supply side factors, high-income households have a greater demand for healthful foods.

The demand for soft drinks and other comparable drinks has been examined in many settings as policymakers have been considering the impacts of "soda taxes" (See Allcott et al. (2019) for an extensive overview). Many papers examine how these taxes effects sales at the store level (Silver et al., 2017; Seiler et al., 2019). However, this fails to capture the aggregate effect since there is evidence that sales in neighboring towns increases and consumers move across borders for purchases (Seiler et al., 2019; Bollinger and Sexton, 2018).

Other research has attempted to find the aggregate effect of SSB taxes on each consumer. Sturm et al. (2010) examines soda taxes in a broad sense rather than a specific SSB tax to see which types of consumers are effected by them the most. Falbe et al. (2016) examined the SSB taxes in San Francisco, Oakland and Berkeley by using repeated cross-sectional surveys to estimate how much soft drink consumption decreased across consumers. Cawley et al. (2018) finds that consumers in Boulder purchased 8.9 ounces of SSB less per shopping trip.

Allcott et al. (2018) used the obesity costs outlined above and estimates the external as well as internal costs of SSB consumption. They use these estimates to propose an optimal tax rate for the US between 1 and 2.1 cents per ounce, while the optimal tax rate for a city is between 0.5 and 1 cent per ounce because of the availability of cross-border shopping.

Our analysis differs from the above analysis in that it examines whether the households we are looking at purchase soft drinks at all, rather than examining how much soft drinks each household consumes: we are focused at looking at the extensive effect of the tax. Thus, our approach will look at how effective the Cook County SSB tax was at decreasing the fraction of families that consumed soft drinks. This could be a particularly important treatment effect since sugar can be addictive (see Avena et al. (2008) for an overview) so eliminating it from the household could be a more effective long-term outcome.

### 3.5.4 Results

#### 3.5.4.1 Model

We want to compare the soda consumption between the Cook County (Chicago) and neighboring suburbs before and after the SSB tax was implemented. The basic linear graphical depiction of the model is shown below in Figure 3.3.

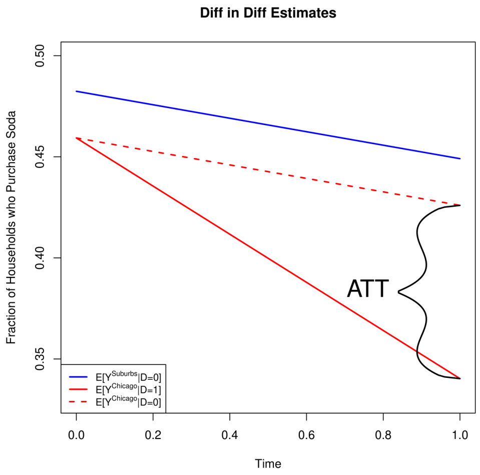

Figure 3.3: This is a graphical depiction of the linear estimation of the ATT.

Our model follows similarly to the example used in Section 3.2.1. $D = 1$ denotes households who live in Cook County while $D = 0$ denotes households who live in neighboring suburbs. At both time points, households make decisions on whether they will purchase ($Y_t = 1$) soda or not ($Y_t = 0$) with a "price" of $Z_t$. $B_1$ represents the willingness to pay for soda and $B_3$ represents the treatment effect. $V_2$ summarized both trends between time periods in demand for soda as well as asecond period shocks. The latent outcome variable, $Y_t^*$, can be interpreted as the "price" minus the household willingness to pay, or their utility from purchasing soda[7].

[7] Note that because we are doing price deviations, $B_1$ and $B_3$ is their meaning (willingness to pay and treatment effect receptively) minus the average price faced by the consumer. This

Using basic ”price” as a special regressor can be particularly dangerous. Price is likely correlated with a consumer’s willingness to pay since consumers with higher willingness to pay will go to stores where goods are more expensive. Remember based on Assumptions A.3.3.1 and A.3.4.1, our special regressor must be independent of $B_1, B_3$, and $V_2$. To avoid this problem, we will focus on price deviations, $\bar{P}_i - P_{it}$, which will take this effect out. We estimate $\bar{P}_i$ by averaging out all $P_{it}$, SL price indicies, across time. For ease of notation, these price deviations will be labeled as $P$.

Note that Assumption A.3.5 is most likely violated because we do not have enough price variation to identify the entire CDFs that are required with the basic CHW estimator. For this reason, we will use our semiparametric estimator, CHW-P, as outlined in Extension 3.

There may be additional concern about price endogeneity. We will generate a control function using nonparametric regression to predict $P_t$ based on $P_{t-1}$, thus using $P_{t-1}$ as my instrumental variable, and calculate $\widehat{\epsilon_i}$ as the residuals from these estimates. We can examine whether it is necessary to include $\widehat{\epsilon_i}$ in my estimates by regressing quantity of $P_t$ and $\widehat{\epsilon_i}$ and testing whether $\widehat{\epsilon_i}$ has a significant effect. We find that the p-value of excluding $\widehat{\epsilon_i}$ as calculated using the control function is 0.77, implying that it might not be necessary. Further, we find that the percentage of soda products on sale and on display is about the same in both time periods. However, we will implement our estimator to compare the results when we control for endogeneity versus when we do not. We will use both Extensions 3 and 4 by conditioning our estimation on $\widehat{\epsilon_i}$ and call this estimate our CHW-P-IV estimator.

#### 3.5.4.2 Numerical Results

Below in Table 3.2 we have the different estimates of ATT as well as bootstrap estimates of the standard error and confidence bands. We can conclude that the

has a scale effect on utility but not impact of our estimates of CATT or ATT.

average treatment effect of the SSB tax in Cook County lead to 9.5-9.8% drop in probability each household that was treated consumes soda. Notice that both the estimates using CHW-P and CHW-P-IV are larger than the LPM and Probit estimation. Thus, using more basic estimators will cause you to underestimate the effect of the SSB tax even though the true average effect is contained in the confidence bounds.

Table 3.2: Empirical ATT Estimates

| | CHW-P | CHW-P-IV | LPM | Probit |
|---|---|---|---|---|
| ATT | -0.0949 | -0.0978 | -0.0856 | -0.0862 |
| SE of ATT | (0.0200) | (0.0678) | (0.0211) | (0.0206) |
| 90% Bands | [-0.114,-0.057] | [-0.274, -0.030] | [-0.119, -0.060] | [-0.119, -0.055] |

This table contains our estimates of the ATT in the first row. Bootstrap estimates of the standard error and 90% confidence bands are included in rows two and three respectively

Because we made additional parametric assumptions, we can learn additional information that is unattainable using more basic methods. Beyond the ATT that we have found, we have the distribution of the CATT conditioned on $P$, as shown in Figure 3.4. The standard deviation of the CATT effects for CHW-P was 0.0397, while for CHW-P-IV was larger at 0.0576 which might just come from a smaller sample size and a decreased range of the support of $P$ because of our conditioning. It also might come from the curse of dimensionality by conditioning on another variable. Here you can see that the effect is significantly right skewed for both estimators, which meant that while the effect was large for most people, about 10% of the population has an effect greater than zero. The quantiles of the distribution are shown in Table 3.3.

Table 3.3: Empirical Quantile Estimates

| | 0% | 20% | 40% | 60% | 80% | 100% |
|---|---|---|---|---|---|---|
| CHW-P | -0.1618 | -0.1216 | -0.1049 | -0.0920 | -0.0749 | 0.0814 |
| CHW-P-IV | -0.3864 | -0.1330 | -0.1117 | -0.1017 | -0.0631 | 0.0813 |

This table contains our quantile estimates of the CATT for our CHW-P and CHW-P-IV estimates.

Furthermore, with our estimates we can look at $\hat{F}_{B_1+V_2+B_3|D}(z_2;1)$, which

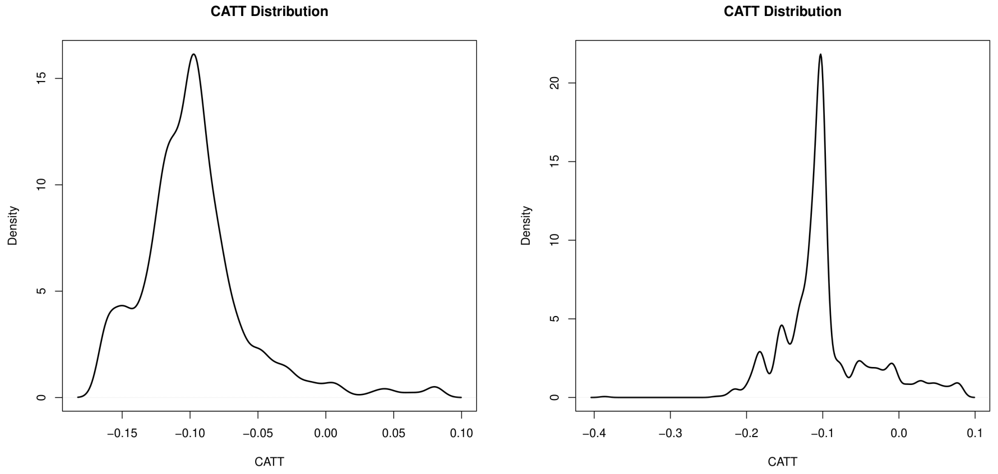

Figure 3.4: Density of the CATT conditional on the special regressor, price deviation. The one on the left is our CHW-P estimator CATT while the one on the right is our CHW-P-IV estimator of CATT.

is just $\widehat{P}[Y_2 = 1|D = 1, Z_2 = z_2]$ and $\hat{F}_{B_1+V_2+B_3|D}(z_2; 1)$ which is $\widehat{P}[Y_2 = 1|D = 0, Z_2 = z_2]$. Our CHW-P estimates are shown in Figure 3.5. From examining these distributions you can see that generally, the higher he likelihood that the consumer purchases soda without the tax, the less likely the tax is to have en effect. This implies that this tax has no effect on the individuals most likely to purchase soda.

Note that we also have the distributions of our other random coefficients, which can provide other insights depending on the setting. Because our $B_1$ distribution is not centered on 1, it is clear that our households respond differently to the tax than to a similar price change.

Examining the CHW-P-IV estimates are trickier to visualize because our CATT estimates are conditioned on $\widehat{\epsilon}$ as well as $P$. I will first plot the CDF of $\hat{F}_{B_1+V_2+B_3|D,\epsilon}(z_2; 1)$ and $\hat{F}_{B_1+V_2|D,\epsilon}(z_2; 1)$ in Figure 3.6. Note that they have a similar shape: The likelihood to consume soda at large negative price deviations is high but never increased about 0.5 in the data. The effect of positive price deviations differ between individuals that have high $\widehat{\epsilon}$ and low $\widehat{\epsilon}$ values. Individuals with high $\widehat{\epsilon}$ values were much more likely ton consume soda no matter

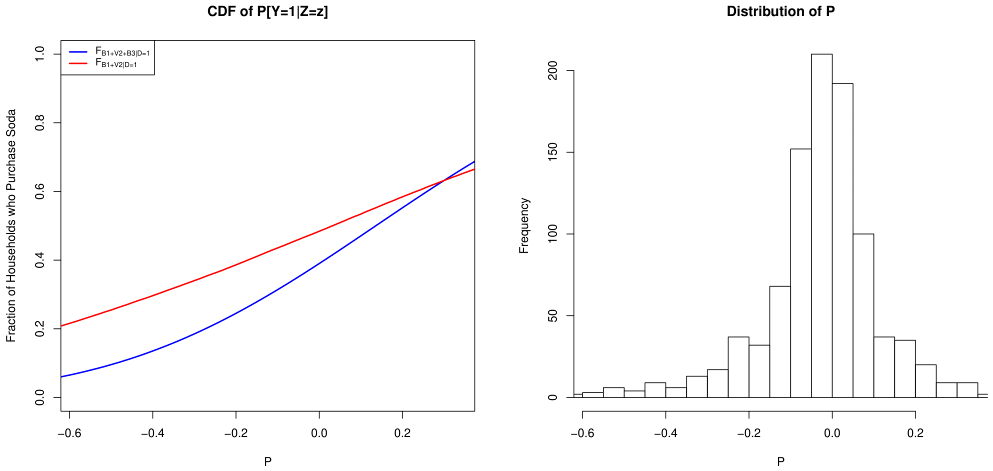

Figure 3.5: These graphs show the CDFs for our CHW-P estimates. Graph on the left shows our estimates for $\hat{F}_{B_1+V_2+B_3|D}(z_2;1)$ and $\hat{F}_{B_1+V_2|D}(z_2;1)$. The graph on the right is a histogram to understand the density of the distribution of price-deviations, $P$.

what the price deviation was. This makes sense since these households generally either faced higher prices the period before or lower prices in the current period.

Our CATT estimates are taken by taking the difference between the two distributions shown in Figure 3.6. This distribution, as well as the joint density of our data, is shown in Figure 3.7. From here we can see that the effect was the the smallest for the individuals that faced the lowest prices (highest price deviations), while it was largest on individuals who did not seem to have any price deviation, but had low $\widehat{\epsilon}$ or those who had a $\widehat{\epsilon}$ near zero and low price deviations. This would imply it had the greatest effect on individuals who had faced low prices in the period before, but had their prices return to normal. This implies that the treatment might be effective at preventing households from building a habit of purchasing soda.

By constructing the CATT distribution, we can focus on specific subgroups of individuals and their ATT. We will focus on two groups of individuals that have been a focus in the literature: households with children and low-income households (Dubois et al., 2019; Han and Powell, 2013; Drewnowski

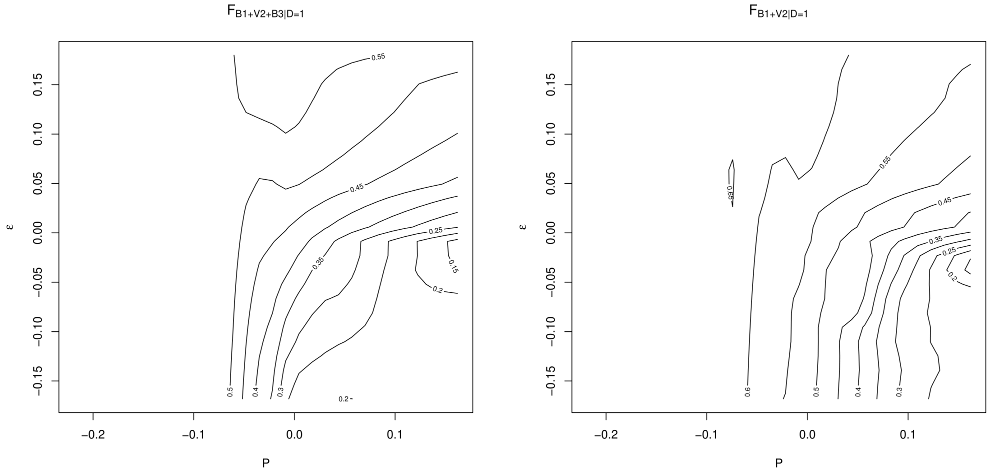

Figure 3.6: These graphs show the CDFs for our CHW-P-IV estimates. Graph on the left shows our estimates for $\hat{F}_{B_1+V_2+B_3|D,\epsilon}(z_2;1)$ while the graph on the right shows our estimates for $\hat{F}_{B_1+V_2|D,\epsilon}(z_2;1)$.

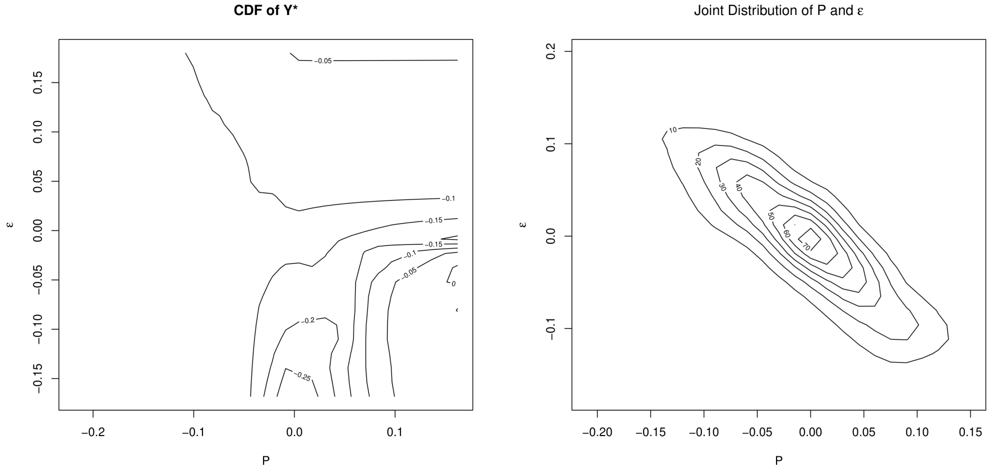

Figure 3.7: These graphs show distribution of CATT conditional on $\widehat{\epsilon}$ and $P$ for our CHW-P-IV estimates. Graph on the left shows our estimates of CATT. The graph on the right shows the joint distribution of price-deviations, $P$, and $\widehat{\epsilon}$.

and Specter, 2004; Currie, 2009). We define low-income households as households with lower than $35,000 of yearly income. The ATT for the households with low-income was -0.0942 and for households with children was -0.0910. Based on the CATT distribution of each of these subgroups, shown below in Figure 3.8, this slight decrease in ATT for the households with children come from an increase density of the tail. Thus, a slightly higher fraction of these households have an ATT close to zero compared to the whole population. Low-income households CATT distribution is very similar to the population CATT distribution.

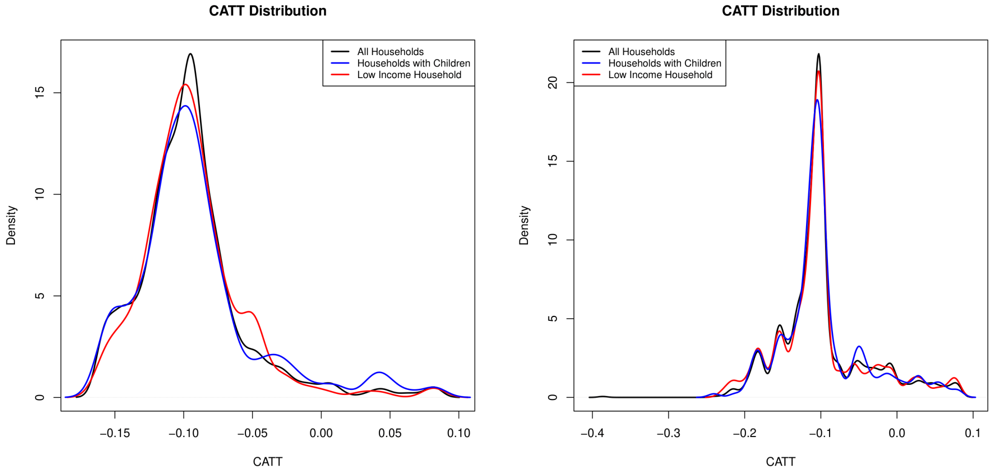

Figure 3.8: These graphs compare the CATT distribution for the total group of households and the subgroups of households with low income and households with children.

Note that across these results, our CHW-P-IV estimates follow our CHW-P results closely. The CHW-P-IV is slightly larger so using the CHW-P results may be underestimating the bias but this difference is minor. Note that the spread of the CHW-P-IV estimator is much larger and may come from the curse of dimensionality. However, if we were to use price rather than price deviation, our results change significantly and we get a ATT of about -0.25. This shows evidence that using price alone would have endogeneity but using price deviation appears to control for the endogeneity.

### 3.5.5 Implications

Note that both the CHW-P and CHW-P-IV estimates show a similar pattern in which consumers generally are ten percent less likely to consume soda after the soda tax but there remains a group between 5-10% of the population for which the tax has no effect. We were able to show that these individuals are the same individuals that were most likely to consume soda without the soda tax.

This is all evidence that is consistent with the idea that consumers who consume the most soda are addicted and cannot be persuaded to stop consuming soda by a relatively small soda tax. This coincides with the research by Allcott et al. (2019) and Bernheim and Rangel (2004) which find similar patterns of addiction or habit formation in terms of responding to taxes for “sin goods” like soda and cigarettes. This might encourage a more focused policy on preventing teenagers from consuming soda to prevent such addiction suggested by O’Donoghue and Rabin (2003). However, we do not find that our treatment has a significantly larger or different effect on households with teenagers or children (see Figure 3.8) which is predicted by O’Donoghue and Rabin (2003).

## 3.6 Summary and Conclusions

This paper is the first to extend the Difference-in-Differences methodology to binary outcome variables. Additionally, our nonparametric random coefficient representation rids us of functional form assumptions and allows for unobserved heterogeneity to be correlated with treatment status. We show identification of the average treatment effect on the treated (ATT) with and without covariates and as a further extension, the identification of the full joint distribution of the latent outcome variables. We propose a sample counterpart estimator for the ATT that we evaluate with the help of Monte Carlo simulations. These show favorable small sample properties compared to the estimators conventionally used for binary choice Difference-in-Differences estimation.

We provide an empirical application to estimate the effect of a Sugar-Sweetened Beverages tax on customer's likelihood to consume soft drinks. We find that the tax in Cook County, Illinois led to consumers on average to decrease their likelihood of consuming soft drinks by about 10% while about 10% of consumers (who were most likely to consume soft drinks previously) did not change their likelihood to consume soft drinks. This is consistent with the previous literature that some consumers are addicted to "sin goods".

Further research could focus on extending the model to discrete outcome variables with more than two outcomes. We also leave the assessment of large-sample properties of our proposed estimator to future research. Further research could extend our estimator to a panel data framework or with a continuous treatment effect. Examples of possible applications of our estimator were outlined in the Introduction and include estimating the effect of job training on unemployment, tax incentives on health insurance, and public policy that relies on individual's willingness to pay.

## 3.7 Appendix

### 3.7.1 Mathematical Appendix

**Proof of Theorem 3.1**: Our main object of interest is the ATT. We can rewrite the ATT in the following way:

$$ATT = \int CATT(z_2) \quad F_{Z_2}(dz_2)$$

The $CATT(z_2)$ can then be rewritten as:

$$\begin{aligned}
CATT(z_2) &= P^1[Y_2 = 1|D = 1, Z_2 = z_2] - P^0[Y_2 = 1|D = 1, Z_2 = z_2] \\
&= P(B_1 + V_2 + B_3 < z_2|D = 1, Z_2 = z_2) - P(B_1 + V_2 < z_2|D = 1, Z_2 = z_2) \\
&= P(B_1 + V_2 + B_3 < z_2|D = 1) - P(B_1 + V_2 < z_2|D = 1) \\
&= F_{B_1+V_2+B_3|D}(z_2; 1) - F_{B_1+V_2|D}(z_2; 1)
\end{aligned}$$

The first line writes out the definition of the CATT in the context of a binary outcome as the difference between the actual probability of success of a treated individual and the counterfactual probability of success had a treated individual not been treated, both conditional on $z_2$. The equality between line two and line three makes use of our assumptions **A.3.3** and **A.3.4** that yield $B_1, B_3, V_2 \perp Z|D$.

We need: $F_{B_1+V_2+B_3|D}(z_2; 1)$ and $F_{B_1+V_2|D}(z_2; 1)$.

Obtaining $F_{B_1+V_2+B_3|D}(z_2; 1)$ is straightforward, as this part is directly observable in the data:

$$\begin{aligned} F_{B_1+V_2+B_3|D}(z_2;1) &= P[B_1+V_2+B_3<z_2|D=1] \\ &= P[Y_2=1|D=1,Z_2=z_2] \end{aligned}$$

Obtaining $F_{B_1+V_2|D}(z_2;1)$ is more involved, as this part is counterfactual and thus not directly observable in the data:

$$\begin{aligned} F_{B_1+V_2|D}(z_2;1) &= \int_{-\infty}^{z_2} f_{B_1+V_2|D}(t;1)dt \\ &= \int_{-\infty}^{z_2}\int_{-\infty}^{\infty} f_{V_2|D}(y;1)f_{B_1|D}(t-y;1)dydt \\ &= \int_{-\infty}^{z_2}\int_{-\infty}^{\infty} f_{V_2|D}(y;0)f_{B_1|D}(t-y;1)dydt \\ &= \int_{-\infty}^{z_2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{V_2,B_1|D}(y,x;0)dxf_{B_1|D}(t-y;1)dydt \\ &= \int_{-\infty}^{z_2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{B_1+V_2,B_1|D}(y+x,x;0)dxf_{B_1|D}(t-y;1)dydt \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{B_1+V_2,B_1|D}(y+x,x;0)dx\int_{-\infty}^{z_2} f_{B_1|D}(t-y;1)dtdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{B_1+V_2,B_1|D}(y+x,x;0)dxF_{B_1|D}(z_2-y;1)dy \end{aligned}$$

The first line uses the definition of the CDF. The second and third line make use of assumption **A.3.3**. In the second line we use the convolution formula for independent random variables as $V_2 \perp B_1$. In the third line we use $V_2 \perp D$. The fourth line reformulates the conditional density of $V_2$ as the marginal density of the joint conditional density of $V_2$ and $B_1$. The fifth line uses the change of variables formula. The sixth line pulls in the limits of integration for $dt$ and the seventh line again uses the definition of the CDF.

Given assumption **A.3.5**, $f_{B_1+V_2,B_1|D=0}$ and $F_{B_1|D=1}$ are directly observable in the data, where

$$f_{B_1+V_2,B_1|D=0} = \partial_{z_1}\partial_{z_2}P[Y_2 = 1, Y_1 = 1|D = 0, Z = z],$$

$$F_{B_1|D=1} = P[Y_1 = 1|D = 1, Z_1 = z_1].$$

This completes the proof. ■

**Alternative Proof of Theorem 3.1 going only through densities:**

We can rewrite the ATT in the following way:

$$ATT = \int CATT(z_2) \quad F_{Z_2}(dz_2)$$

The $CATT(z_2)$ can then be rewritten as:

$$\begin{aligned}
CATT(z_2) &= P^1[Y_2 = 1|D = 1, Z_2 = z_2] - P^0[Y_2 = 1|D = 1, Z_2 = z_2] \\
&= P(B_1 + V_2 + B_3 < z_2|D = 1, Z_2 = z_2) - P(B_1 + V_2 < z_2|D = 1, Z_2 = z_2) \\
&= P(B_1 + V_2 + B_3 < z_2|D = 1) - P(B_1 + V_2 < z_2|D = 1) \\
&= F_{B_1+V_2+B_3|D}(z_2; 1) - F_{B_1+V_2|D}(z_2; 1) \\
&= \int_{-\infty}^{z_2} f_{B_1+V_2+B_3|D}(t; 1)dt - \int_{-\infty}^{z_2} f_{B_1+V_2|D}(t; 1)dt \\
&= \int_{-\infty}^{z_2} [f_{B_1+V_2+B_3|D}(t; 1) - f_{B_1+V_2|D}(t; 1)]dt
\end{aligned}$$

We need: $f_{B_1+V_2+B_3|D}(t; 1)$ and $f_{B_1+V_2|D}(t; 1)$.

Start off with:

$$\begin{aligned}
P[Y_2 = 1, Y_1 = 1|D = 1, Z = z] &= P[B_1 + V_2 + B_3 < z_2, B_1 < z_1|D = 1] \\
&= F_{B_1+V_2+B_3,B_1|D}(z_2, z_1; 1) \\
&= \int_{-\infty}^{z_2} \int_{-\infty}^{z_1} f_{B_1+V_2+B_3,B_1|D}(x, y; 1)dydx
\end{aligned}$$

Taking the derivative with respect to $z_2$ and $z_1$ yields:

$$\partial_{z_1}\partial_{z_2} P[Y_2 = 1, Y_1 = 1|D = 1, Z = z] = f_{B_1+V_2+B_3,B_1|D}(z_2, z_1; 1)$$

Obtain the first density via integration:

$$f_{B_1+V_2+B_3|D}(z_2;1) = \int_{-\infty}^{\infty} f_{B_1+V_2+B_3,B_1|D}(z_2,t;1)dt$$

Also via integration obtain:

$$f_{B_1|D}(z_1;1) = \int_{-\infty}^{\infty} f_{B_1+V_2+B_3,B_1|D}(t,z_1;1)dt$$

If we know $f_{V_2}$ we can do convolution and we are done.

Use $P[Y_2 = 1, Y_1 = 1|D = 0, Z = z]$ similarly as above and obtain $f_{B_1+V_2,B_1|D}(z_2, z_1; 0)$.

By change of variables:

$$f_{V_2,B_1|D}(z_2 - z_1, z_1; 0) = f_{B_1+V_2,B_1|D}(z_2, z_1; 0)$$

Obtain $f_{V_2|D}(z_2 - z_1; 0)$ via integration.

$$f_{V_2|D}(z_2 - z_1; 0) = \int_{-\infty}^{\infty} f_{V_2,B_1|D}(z_2 - z_1, t; 0)dt$$

Then use $V_2 \perp D$ and thus $f_{V_2|D}(z_2 - z_1; 0) = f_{V_2}(z_2 - z_1) = f_{V_2|D}(z_2 - z_1; 1)$.

Now use convolution:

$$f_{B_1+V_2|D}(z_2 - z_1; 1) = \int_{-\infty}^{\infty} f_{B_1|D}(t; 1)f_{V_2}((z_2 - z_1) - t)dt$$

This completes the proof. ■

**Proof of Extension 1:**

Define:

$$\begin{aligned}
\psi(z_1, z_2, d) &= \partial_{z_1}\partial_{z_2} P\left[Y_2 = 1, Y_1 = 1 | D = d, Z = z\right] \\
&= \partial_{z_1}\partial_{z_2} \int_{\mathcal{B}} \mathbf{1}\left\{z_2 > b_2 + b_3 d\right\} \mathbf{1}\left\{z_1 > b_1\right\} f_{B|D}(b; d) db \\
&= \partial_{z_1}\partial_{z_2} \int_{-\infty}^{z_2} \int_{-\infty}^{z_1} \int_{\mathcal{B}} \mathbf{1}\left\{s_2 = b_2 + b_3 d\right\} \mathbf{1}\left\{s_1 = b_1\right\} f_{B|D}(b; d) db ds \\
&= \int_{\mathcal{B}} \mathbf{1}\left\{z_2 = b_2 + b_3 d\right\} \mathbf{1}\left\{z_1 = b_1\right\} f_{B|D}(b; d) db
\end{aligned}$$

The equality between line one and line two makes use of our assumptions **A.3.3** and **A.3.4** which yield $B_1, B_3, V_2 \perp Z|D$.

Now consider:

$$\begin{aligned}
& E\left\{ \left. \frac{\exp\left[i(sZ_1 + tZ_2)\right] \psi(Z_1, Z_2, d)}{f_{Z|D}(Z; d)} \right| D = d \right\} \\
= \; & \int_{\mathcal{Z}} \int_{\mathcal{B}} \exp\left[i(sz_1 + tz_2)\right] \mathbf{1}\left\{z_2 = b_2 + b_3 d\right\} \mathbf{1}\left\{z_1 = b_1\right\} f_{B|D}(b; d) db dz \\
= \; & E\left\{\exp\left[i(sB_1 + t(B_2 + B_3 d))\right] | D = d\right\} \\
= \; & E\left\{\exp\left[i((s + t)B_1 + t(V_2 + B_3 d))\right] | D = d\right\}
\end{aligned}$$

Let $\Delta Z = Z_1 - Z_0$. Hence, with $s = -t$, and for $d = 0$, we get:

$$E\left\{ \left. \frac{\exp(it\Delta Z)\psi(Z_1, Z_2, 0)}{f_{Z|D}(Z; 0)} \right| D = 0 \right\} = E\left\{ \exp(itV_2) \right| D = 0\}$$

And with $s + t = \sigma$, and for $d = 1$, we get:

$$E\left\{ \left. \frac{\exp\left[i(\sigma Z_1 + t(Z_2 - Z_1))\right] \psi(Z_1, Z_2, 1)}{f_{Z|D}(Z; 1)} \right| D = 1 \right\} = E\left\{\exp\left[i(\sigma B_1 + t(V_2 + B_3))\right] | D = 1\right\}$$

Under the assumption that $V_2$ is independent of $B_3, B_1|D = 1$, we get:

$$E\left\{\exp\left[i(\sigma B_1 + tB_3)\right]|D = 1\right\} E\left\{\exp(itV_2)|D = 1\right\}$$

And thus, as $E\left\{\exp(itV_2)|D = 1\right\} = E\left\{\exp(itV_2)|D = 0\right\}$, due to assumption **A.3.3**, we obtain:

$$\frac{E\left\{\left.\frac{\exp[i(\sigma Z_1 + t(Z_2 - Z_1))]\psi(Z_1, Z_2, 1)}{f_{Z|D}(Z;d)}\right| D = 1\right\}}{E\left\{\left.\frac{\exp(it\Delta Z)\psi(Z_1, Z_2, 0)}{f_{Z|D}(Z|0)}\right| D = 0\right\}} = E\left\{\exp\left[i(\sigma B_1 + tB_3)\right]|D = 1\right\},$$

This is the joint characteristic function of $B_1$ and $B_3$ conditional on $D = 1$. The joint characteristic function of $B_1$, $B_3$, and $V_2$ conditional on $D = 1$ can then easily be constructed in the following way:

$$\begin{aligned}
\Phi_{B_1,B_3,V_2|D=1} &= E\left\{\exp\left[i(\sigma B_1 + tB_3 + rV_2)\right]|D = 1\right\} \\
&= E\left\{\exp\left[i(\sigma B_1 + tB_3)\right]|D = 1\right\} \times E\left\{\exp(irV_2)|D = 1\right\} \\
&= E\left\{\exp\left[i(\sigma B_1 + tB_3)\right]|D = 1\right\} \times E\left\{\exp(irV_2)|D = 0\right\} \\
&= \frac{E\left\{\left.\frac{\exp[i(\sigma Z_1 + t(Z_2 - Z_1))]\psi(Z_1, Z_2, 1)}{f_{Z|D}(Z;1)}\right| D = 1\right\}}{E\left\{\left.\frac{\exp[it(Z_2 - Z_1)]\psi(Z_1, Z_2, 0)}{f_{Z|D}(Z;0)}\right| D = 0\right\}} \times \\
&\quad E\left\{\left.\frac{\exp\left[ir(Z_2 - Z_1)\right]\psi(Z_1, Z_2, 0)}{f_{Z|D}(Z;0)}\right| D = 0\right\}
\end{aligned}$$

This completes the proof. ■

### 3.7.2 Understanding Restrictions on Coefficients of $Z_t$

**Original Model:**

$$\begin{aligned} \tilde{Y_1^*} &= \tilde{B_1} - \Gamma Z_1 \\ \tilde{Y_2^*} &= \tilde{B_2} + \tilde{B_3} D - \Gamma Z_2 \\ \tilde{B_2} &= \tilde{B_1} + \tilde{V_2} \\ Y_t &= \mathbf{1}\{\tilde{Y_t^*} < 0\} \end{aligned}$$

**Restrictions:**

1. The random coefficients on $Z_1$ and $Z_2$ are the same, i.e. the effect that $Z_t$ has on the latent variable $Y_t^*$ stays constant over the two time periods $t = 1, 2$.

2. $\Gamma \neq 0$ and we know the sign of $\Gamma$, i.e. there is an effect of $Z_t$ on $Y_t^*$ and we know its direction.

For $\Gamma > 0$, we can divide the first two equations by $\Gamma$ and obtain:

$$\begin{aligned} \frac{\tilde{Y_1^*}}{\Gamma} &= \frac{\tilde{B_1}}{\Gamma} - Z_1 \\ \frac{\tilde{Y_2^*}}{\Gamma} &= \frac{\tilde{B_2}}{\Gamma} + \frac{\tilde{B_3}}{\Gamma} D - Z_2 \\ \tilde{B_2} &= \tilde{B_1} + \tilde{V_2} \\ Y_t &= \mathbf{1}\{\tilde{Y_t^*} < 0\} \end{aligned}$$

**Remark B.1:** This last assumption is not restrictive, as we can always divide $Z_t$'s by minus one to obtain a positive $\Gamma$.

Now let's redefine our coefficients: $Y_1^* = \frac{\tilde{Y_1^*}}{\Gamma}, Y_2^* = \frac{\tilde{Y_2^*}}{\Gamma}, B_1 = \frac{\tilde{B_1}}{\Gamma}, B_2 = \frac{\tilde{B_2}}{\Gamma}, B_3 = \frac{\tilde{B_3}}{\Gamma}, V_2 = \frac{\tilde{V_2}}{\Gamma}$.

This leaves us with the basic model introduced in the main text.

$$
\begin{aligned}
Y_1^* &= B_1 - Z_1 \\
Y_2^* &= B_2 + B_3 D - Z_2 \\
B_2 &= B_1 + V_2 \\
Y_t &= \mathbf{1}\{Y_t^* < 0\}
\end{aligned}
$$

### 3.7.3 Data

There are a few issues to keep in mind when dealing with this Homescan data. The first issue is with misreporting of quantity. Einav et al. (2010) examines which goods are more likely to be subject to this error. They find that consumable goods like small drinks (like many soft drinks) is likely to be consumed before getting home so are more likely to not be scanned. This could add noise to our results, but should not bias the results because quantity is only a dependent variable and we assume that these problems effect those in Cook County as well as those in the suburbs equally.

Another source of measurement error that is more concerning can come from the price. Individuals record their purchases by scanning the items they buy when they get home. The individuals input the quantity they purchase and Nielsen matches it with the average price of the good at the store where they purchased it that week. This can lead to two types of errors. The first comes from the price changing in the middle of the week. These types of errors are approximately normally distributed.

The second type of error comes from not including discounts from loyalty cards. Einav et al. (2010) examines a retailer used in the Homescan data which has loyalty cards and finds that loyalty cards are used in about 75-80% of the transactions. Further, this would bias my prices upwards, which when comparing Homescan data with data from the retailer finds that the prices used in the Homescan data is about 7% higher. On the other hand, these price measurement errors may be overestimated since some retailers do not have loyalty cards at all. Further, Homescan data errors are comparable to errors found in other commonly used data sets Einav et al. (2010); Aguiar and Hurst (2007); Lin (2018). Additional examination of this measurement error and it's effect on the results is left for future research.

When there is no good purchased, we find the average price plus tax for each month at the store the most commonly purchase soft drinks at by matching with Nielsen Retail Scanner Data. If I am unable to identify the store where the individual commonly purchases soft drinks in the month, of if the store's prices are unavailable, I estimate the prices based on average prices at stores used by similar consumers. The subset of similar consumers I choose is explained below. If there were no prices in the subset I tried to match, I moved to a broader subset below.

1. Individuals with the same favorite retail chain, income level, county and zip-code
2. Individuals with the same favorite retail chain, income level, and county
3. Individuals with the same favorite retail chain, and zip-code
4. Individuals with the same favorite retail chain, and county
5. Individuals with the same favorite retail chain, income level, whether they lived in Cook County, and designated market
6. Individuals with the same favorite retail chain, whether they lived in Cook County, and designated market
7. Individuals with the same zip-code
8. Individuals with the same favorite retail chain, and designated market
9. Individuals with the same income level and county
10. Individuals with the same income level and whether they lived in Cook

County

11. Individuals with the same county
12. Individuals on whether they lived in Cook County
13. Individuals with the same designated market
14. Individuals with the same favorite retail chain
15. All individuals

Below is the Summary Statistics for the Nielsen Scanner Data we used in application for each month and the total data across both months. Our sample contains 1,008 households in Cook County and 1,080 households from neighboring counties.

Table 3.4

| | July 2017 | September 2017 | Total |
|---|---|---|---|
| SL Price Index | 0.5639 | 0.5886 | 0.5762 |
| | (0.4153) | (0.4067) | (0.4112) |
| Price Deviation | -0.0059 | 0.0188 | 0.0064 |
| | (0.2634) | (0.2391) | (0.2518) |
| $\epsilon_i$ | -0.0155 | 0.0080 | -0.0038 |
| | (0.2451) | (0.2195) | (0.2329) |
| Quantity | 0.4713 | 0.3966 | 0.4339 |
| Chicago | 0.4828 | 0.4828 | 0.4828 |

Estimate of each variable mean for each time period and across both time periods is included. The Standard Deviation is included below in parenthesis. Quantity is a measure of whether the individual purchased soft drinks. Chicago is a measure of whether they lived in Chicago.$\epsilon_i$ is the errors left over from the control functions described previously.

# Bibliography

Aasnass, J. and A. Rødseth (1983). Engel curves and systems of demand functions. *European Economic Review 20*(1), 95–121.

Abadie, A. and S. Dermisi (2008). Is terrorism eroding agglomeration economies in central business districts? Lessons from the office real estate market in downtown Chicago. *Journal of Urban Economics 64*(2), 451–463.

Abaluck, J. and A. Adams (2017). What do consumers consider before they choose? identification from asymmetric demand responses. Technical report, National Bureau of Economic Research.

Aguiar, M. and E. Hurst (2007). Life-cycle prices and production. *American Economic Review 97*(5), 1533–1559.

Allcott, H., R. Diamond, J.-P. Dubé, et al. (2017). The geography of poverty and nutrition: Food deserts and food choices across the united states.

Allcott, H., B. B. Lockwood, and D. Taubinsky (2018). Regressive sin taxes, with an application to the optimal soda tax. *Unpublished. https://sites. google. com/site/allcott/research.*

Allcott, H., B. B. Lockwood, and D. Taubinsky (2019). Should we tax sugar sweetened beverages? an overview of theory and evidence. *Unpublished. https://sites. google. com/site/allcott/research.*

Altonji, J. G. and R. L. Matzkin (2005). Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica 73*(4), 1053–1102.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society: Series B (Methodological) 32*(2), 283–301.

Ando, T. and J. Bai (2016). Panel data models with grouped factor structure under unknown group membership. *Journal of Applied Econometrics 31*(1), 163–191.

Arellano, M. (2003). Discrete choices with panel data. *Investigaciones Economicas 27*, 423–458.

Arellano, M. and S. Bond (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The review of economic studies 58*(2), 277–297.

Arellano, M. and S. Bonhomme (2012). Identifying distributional characteristics in random coefficients panel data models. *The Review of Economic Studies 79*(3), 987–1020.

Asano, S. and E. P. Fiuza (2015). Estimation of the brazilian consumer demand system.

Ashenfelter, O. (1978). Estimating the effect of training programs on earnings. *The Review of Economics and Statistics 60*(1), 47–57.

Ashenfelter, O. and D. Card (1985). Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs. *The Review of Economics and Statistics 67*(4), 648–660.

Athey, S. and G. W. Imbens (2006). Identification and inference in nonlinear difference-in-difference models. *Econometrica 74*(2), 431–497.

Avena, N. M., P. Rada, and B. G. Hoebel (2008). Evidence for sugar addiction: behavioral and neurochemical effects of intermittent, excessive sugar intake. *Neuroscience & Biobehavioral Reviews 32*(1), 20–39.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica 77*(4), 1229–1279.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221.

Baltagi, B. H., J. M. Griffin, and W. Xiong (2000). To pool or not to pool: Homogeneous versus heterogeneous estimators applied to cigarette demand. *Review of Economics and Statistics 82*(1), 117–126.

Banks, J., R. Blundell, and A. Lewbel (1997). Quadratic engel curves and consumer demand. *Review of Economics and statistics 79*(4), 527–539.

Bargain, O., L. González, C. Keane, and B. Özcan (2012). Female labor supply and divorce: New evidence from Ireland. *European Economic Review 56*(8), 1675–1691.

Barseghyan, L., M. Coughlin, F. Molinari, and J. C. Teitelbaum (2019). Heterogeneous choice sets and preferences. *arXiv preprint arXiv:1907.02337*.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica 80*(6), 2369–2429.

Belloni, A. and V. Chernozhukov (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli 19*(2), 521–547.

Belloni, A., V. Chernozhukov, and Y. Wei (2016). Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics 34*(4), 606–619.

Bernheim, B. D. and A. Rangel (2004). Addiction and cue-triggered decision processes. *American economic review 94*(5), 1558–1590.

Berry, S., A. Khwaja, V. Kumar, A. Musalem, K. C. Wilbur, G. Allenby, B. Anand, P. Chintagunta, W. M. Hanemann, P. Jeziorski, et al. (2014). Structural models of complementary choices. *Marketing Letters 25*(3), 245–256.

Berry, S., J. Levinsohn, and A. Pakes (1995a). Automobile prices in market equilibrium. *Econometrica 63*(4), 841–890.

Berry, S., J. Levinsohn, and A. Pakes (1995b). Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, 841–890.

Berry, S. T. and P. A. Haile (2010). Nonparametric Identification of Multinomial Choice Demand Models with Heterogeneous Consumers. *Cowles Foundation Discussion Paper*.

Bester, C. A. and C. B. Hansen (2016). Grouped effects estimators in fixed effects models. *Journal of Econometrics 190*(1), 197–208.

Bickel, P. J., Y. Ritov, A. B. Tsybakov, et al. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics 37*(4), 1705–1732.

Billingsley, P. (1995). Probability and measure. *A Wiley-Interscience Publication, John Wiley*.

Binkley, J. K. and A. Golub (2011). Consumer demand for nutrition versus taste in four major food categories. *Agricultural Economics 42*(1), 65–74.

Blaylock, J., D. Smallwood, K. Kassel, J. Variyam, and L. Aldrich (1999). Economics, food choices, and nutrition. *Food Policy 24*(2-3), 269–286.

Blundell, R. and J.-M. Robin (2000). Latent separability: Grouping goods without weak separability. *Econometrica 68*(1), 53–84.

Bochner, S., K. Chandrasekharan, et al. (1949). *Fourier transforms.* Number 19. Princeton University Press.

Bollinger, B. and S. Sexton (2018). Local excise taxes, sticky prices, and spillovers: Evidence from berkeley's soda tax. *Sticky Prices, and Spillovers: Evidence from Berkeley's Soda Tax (January 12, 2018).*

Bonhomme, S., T. Lamadon, and E. Manresa (2017). Discretizing unobserved heterogeneity. *University of Chicago, Becker Friedman Institute for Economics Working Paper* (2019-16).

Bonhomme, S. and E. Manresa (2015). Grouped patterns of heterogeneity in panel data. *Econometrica 83*(3), 1147–1184.

Bonhomme, S. and U. Sauder (2011). Recovering distributions in difference-in-differences models: A comparison of selective and comprehensive schooling. *Review of Economics and Statistics 93*(2), 479–494.

Browning, M., J. Carro, et al. (2007). Heterogeneity and microeconometrics modeling. *Econometric Society Monographs 43*, 47.

Brusco, M. and D. Steinley (2007). A comparison of heuristic procedures for minimum within-cluster sums of squares partitioning. *Psychometrika 71*, 583–600.

Campolieti, M. and C. Riddell (2012). Disability policy and the labor market: Evidence from a natural experiment in Canada, 1998–2006. *Journal of Public Economics 96*(3), 306–316.

Caplin, A. and M. Dean (2015). Revealed preference, rational inattention, and costly information acquisition. *American Economic Review 105*(7), 2183–2203.

Card, D. (1990). The impact of the Mariel boatlift on the Miami labor market. *Industrial & Labor Relations Review 43*(2), 245–257.

Card, D. and A. B. Krueger (1994). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. *The American Economic Review 84*(4), 772–793.

Carlson, A. and E. Frazão (2012). Are healthy foods really more expensive? it depends on how you measure the price. *It Depends on How You Measure the Price (May 1, 2012). USDA-ERS Economic Information Bulletin* (96).

Cawley, J., D. Frisvold, A. Hill, and D. Jones (2018). The impact of the philadelphia beverage tax on prices and product availability. Working Paper 24990, National Bureau of Economic Research.

Cawley, J. and C. Meyerhoefer (2012). The medical care costs of obesity: an instrumental variables approach. *Journal of health economics 31*(1), 219–230.

Cawley, J., C. Meyerhoefer, A. Biener, M. Hammer, and N. Wintfeld (2015). Savings in medical expenditures associated with reductions in body mass index among us adults with obesity, by diabetes status. *Pharmacoeconomics 33*(7), 707–722.

Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of econometrics 18*(1), 5–46.

Chamberlain, G. (1984). Panel data: in z. griliches and md intriligator. *Handbook of Econometrics 2*.

Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica: Journal of the Econometric Society*, 567–596.

Chamberlain, G. (2010). Binary response models for panel data: Identification and information. *Econometrica 78*(1), 159–168.

Chen, S. E., J. Liu, and J. K. Binkley (2012). An exploration of the relationship between income and eating behavior. *Agricultural and Resource Economics Review 41*(1), 82–91.

Cherchye, L., B. De Rock, R. Griffith, M. O'Connell, K. Smith, and F. Vermeulen (2017). A new year, a new you? heterogeneity and self-control in food purchases.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. K. Newey (2016). Double machine learning for treatment and causal parameters. Technical report, cemmap working paper.

Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2009). Identification and estimation of marginal effects in nonlinear panel models. Technical report, cemmap working paper.

Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2013). Average and quantile effects in nonseparable panel models. *Econometrica 81*(2), 535–580.

Chernozhukov, V., I. Fernandez-Val, S. Hoderlein, H. Holzmann, and W. Newey (2015). Nonparametric identification in panels using quantiles. *Journal of Econometrics 188*(2), 378–392.

Chernozhukov, V., I. Fernández-Val, and W. K. Newey (2019). Nonseparable multinomial choice models in cross-section and panel data. *Journal of Econometrics 211*(1), 104–116.

Chernozhukov, V., C. Hansen, and M. Spindler (2016). High-dimenional metrics in r. *arXiv:1603.01700*.

Chernozhukov, V., J. A. Hausman, and W. K. Newey (2019). Demand analysis with many prices. *Unpublished. https://www.nber.org/papers/w26424*.

Chernozhukov, V., W. Newey, and J. Robins (2018). Double/de-biased machine learning using regularized riesz representers. *arXiv preprint arXiv:1802.08667*.

Chetty, R., A. Looney, and K. Kroft (2009). Salience and taxation: Theory and evidence. *American economic review 99*(4), 1145–77.

Chevalier, J. A., A. K. Kashyap, and P. E. Rossi (2003). Why don't prices rise during periods of peak demand? evidence from scanner data. *The American Economic Review 93*(1), 15–37.

Chintagunta, P. K. and H. S. Nair (2011). Structural workshop paperdiscrete-choice models of consumer demand in marketing. *Marketing Science 30*(6), 977–996.

Chiong, K. X. and M. Shum (2018). Random projection estimation of discrete-choice models with large choice sets. *Management Science 65*(1), 256–271.

Currie, J. (2009). Healthy, wealthy, and wise: Socioeconomic status, poor health in childhood, and human capital development. *Journal of Economic Literature 47*(1), 87–122.

Deaton, A. (1980). An almost ideal demand system. *The American Economic Review 70*(3), 312–326.

Delaigle, A. and A. Meister (2007). Nonparametric regression estimation in the heteroscedastic errors-in-variables problem. *Journal of the American Statistical Association 102*(480), 1416–1426.

Delaigle, A., A. Meister, et al. (2008). Density estimation with heteroscedastic error. *Bernoulli 14*(2), 562–579.

DellaVigna, S. and M. Gentzkow (2017). Uniform pricing in us retail chains. Technical report, National Bureau of Economic Research.

Department of Health (2016). Cut back on sugary drinks.

Dong, Y. and A. Lewbel (2015). A simple estimator for binary choice models with endogenous regressors. *Econometric Reviews 34*(1-2), 82–105.

Drewnowski, A. and N. Darmon (2005). The economics of obesity: dietary energy density and energy cost–. *The American journal of clinical nutrition 82*(1), 265S–273S.

Drewnowski, A. and P. Eichelsdoerfer (2010). Can low-income americans afford a healthy diet? *Nutrition today 44*(6), 246.

Drewnowski, A. and S. E. Specter (2004). Poverty and obesity: the role of energy density and energy costs. *The American journal of clinical nutrition 79*(1), 6–16.

Drukker, D. M. and D. Liu (2019). A plug-in for poisson lasso and a comparison of partialing-out poisson estimators that use different methods for selecting the lasso tuning parameters.

Dubois, P., R. Griffith, and M. O'Connell (2019). How well targeted are soda taxes?

Durlauf, S. N., A. Kourtellos, and A. Minkin (2001). The local solow growth model. *European Economic Review 45*(4-6), 928–940.

Dyer, W. T. and R. W. Fairlie (2004). Do family caps reduce out-of-wedlock births? Evidence from Arkansas, Georgia, Indiana, New Jersey and Virginia. *Population Research and Policy Review 23*(5-6), 441–473.

Einav, L., E. Leibtag, and A. Nevo (2010). Recording discrepancies in nielsen homescan data: Are they present and do they matter? *QME 8*(2), 207–239.

Evdokimov, K. (2010). Identification and estimation of a nonparametric panel data model with unobserved heterogeneity. *Department of Economics, Princeton University*.

Falbe, J., H. Thompson, C. Becker, N. Rojas, C. McCulloch, and K. Madsen (2016). Impact of the berkeley excise tax on sugar-sweetened beverage consumption. *American Journal of Public Health 106*(10), 1865–1871.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association 96*(456), 1348–1360.

Fan, Y. and R. Li (2012). Variable selection in linear mixed effects models. *The Annals of Statistics 40*(4), 2043–2068.

Finkelstein, E. A., J. G. Trogdon, J. W. Cohen, and W. Dietz (2009). Annual medical spending attributable to obesity: Payer-and service-specific estimates: Amid calls for health reform, real cost savings are more likely to be achieved through reducing obesity and related risk factors. *Health affairs 28*(Suppl1), w822–w831.

Forgy, E. (1965). Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications. *Biometrics 21*, 768–769.

Fox, J. T. and A. Gandhi (2016). Nonparametric identification and estimation of random coefficients in multinomial choice models. *The RAND Journal of Economics 47*(1), 118–139.

Gautier, E. and Y. Kitamura (2013). Nonparametric estimation in random coefficients binary choice models. *Econometrica 81*(2), 581–607.

Gentzkow, M. (2007). Valuing new goods in a model with complementarity: Online newspapers. *American Economic Review 97*(3), 713–744.

George, L. and J. Waldfogel (2003). Who affects whom in daily newspaper markets? *Journal of Political Economy 111*(4), 765–784.

Gillen, B., H. Moon, and M. Shum (2014). Demand estimation with high-dimensional produc characteristics. *Advances in Econometrics 34*, 301–323.

Gillen, B. J., S. Montero, H. R. Moon, and M. Shum (2015). Blp-lasso for aggregate discrete choice models of elections with rich demographic covariates. *USC-INET Research Paper* (15-27).

Golan, E. H., H. Stewart, F. Kuchler, and D. Dong (2008). Can low-income americans afford a healthy diet? Technical report.

Graham, B. S. and J. L. Powell (2012). Identification and estimation of average partial effects in irregular correlated random coefficient panel data models. *Econometrica 80*(5), 2105–2152.

Gruber, J. and J. Poterba (1994). Tax incentives and the decision to purchase health insurance: Evidence from the self-employed. *The Quarterly Journal of Economics 109*(3), 701–733.

Hahn, J. and H. Moon (2010). Panel data models with finite number of multiple equilibria. *Econometric Theory 26*(3), 863–881.

Han, E. and L. M. Powell (2013). Consumption patterns of sugar-sweetened beverages in the united states. *Journal of the Academy of Nutrition and Dietetics 113*(1), 43–53.

Hansen, P. and N. Mladenović (2001). J-means: A new local search heuristic for minimum sum-of-squares clustering. *Pattern Recognition 34*(2), 405–413.

Hansen, P., N. Mladenović, and J. A. Moreno Pérez (2010). Automobile prices in market equilibrium. *Annals of Operations Research 175*, 367–407.

Hartman, W. R. (2010). Demand estimation with social interactions and the implications for targeted marketing. *Marketing Science 29*(4), 585–601.

Hauser, J. R. and B. Wernerfelt (1990). An evaluation cost model of consideration sets. *Journal of consumer research 16*(4), 393–408.

Hausman, J. (1996). Valuation of new goods under perfect and imperfect competition. *58*, 207–248.

Hausman, J., G. Leonard, and J. D. Zona (1994). Competitive analysis with differentiated products. *Annales d'Econ. et Stat. 34*, 159–190.

Hausman, J. A., B. H. Hall, and Z. Griliches (1984). Econometric models for count data with an application to the patents-r&d relationship.

Heckman, J. J. and R. Robb Jr (1985). Alternative methods for evaluating the impact of interventions: An overview. *Journal of econometrics 30*(1-2), 239–267.

Hendel, I. and A. Nevo (2006). Measuring the implications of sales and consumer inventory behavior. *Econometrica 74*(6), 1637–1673.

Hitsch, G. J., A. Hortacsu, and X. Lin (2017). Prices and promotions in us retail markets: Evidence from big data.

Hoderlein, S., H. Holzmann, and A. Meister (2017). The triangular model with random coefficients. *Journal of econometrics 201*(1), 144–169.

Hoderlein, S., J. Klemelä, and E. Mammen (2010). Analyzing the random coefficient model nonparametrically. *Econometric Theory 26*(03), 804–837.

Hoderlein, S. and E. Mammen (2007). Identification of marginal effects in nonseparable models without monotonicity. *Econometrica 75*(5), 1513–1518.

Hoderlein, S. and S. Mihaleva (2008). Increasing the price variation in a repeated cross section. *Journal of Econometrics 147*(2), 316–325.

Hoderlein, S. and H. White (2012). Nonparametric identification in nonseparable panel data models with generalized fixed effects. *Journal of Econometrics 168*(2), 300–314.

Honoré, B. E. (1992). Trimmed lad and least squares estimation of truncated and censored regression models with fixed effects. *Econometrica: journal of the Econometric Society*, 533–565.

Honoré, B. E. and A. Lewbel (2002). Semiparametric binary choice panel data models without strictly exogeneous regressors. *Econometrica 70*(5), 2053–2063.

Hsiao, C. and M. H. Pesaran (2008). Random coefficient panel data models. *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*, 187216.

Hsiao, C. and A. K. Tahmiscioglu (1997). A panel analysis of liquidity constraints and firm investment. *Journal of the American Statistical Association 92*(438), 455–465.

Huang, C. L. and B.-H. Lin (2007). A Hedonic Analysis of Fresh Tomato Prices among Regional Markets. *Applied Economic Perspectives and Policy 29*(4), 783–800.

Ichimura, H. (1991). Semiparametric least squares (sls) and weighted sls estimation of single-index models.

Ichimura, H. and T. S. Thompson (1998). Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution. *Journal of Econometrics 86*(2), 269–295.

Imbens, G. W. and W. K. Newey (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica 77*(5), 1481–1512.

Jing, B.-Y., Q.-M. Shao, and Q. Wang (2003). Self-normalized cramr-type large deviations for independent random variables. *Ann. Probab. 31*(4), 2167–2215.

Jorgenson, D. (1990). Aggregate consumer behavior and the measurement of social welfare. *Econometrica 58*(5), 1007–1040.

Kasy, M. (2011). Identification in triangular systems using control functions. *Econometric Theory 27*(3), 663–671.

Keane, M. and K. Wolpin (1997). The career decisions of young men. *Journal of Political Economy 105*(3), 473–522.

Klein, R. W. and R. H. Spady (1993). An efficient semiparametric estimator for binary response models. *Econometrica: Journal of the Econometric Society*, 387–421.

Koulayev, S. (2009). Estimating demand in search markets: The case of online hotel bookings. *FRB of Boston Working Paper* (09-16).

Lechner, M. (2011). The Estimation of Causal Effects by Difference-in-Difference Methods. *Foundations and Trends in Econometrics 4*(3), 165–224.

Lee, K., M. H. Pesaran, and R. Smith (1997). Growth and convergence in a multi-country empirical stochastic solow model. *Journal of applied Econometrics 12*(4), 357–392.

Lewbel, A. (1989). Identification and estimation of equivalence scales under weak separability. *The Review of Economic Studies 56*(2), 311–316.

Lewbel, A. (2000). Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables. *Journal of Econometrics 97*(1), 145–177.

Lewbel, A. (2014). An overview of the special regressor method. *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*, 38–62.

Lewbel, A., D. McFadden, and O. Linton (2011). Estimating features of a distribution from binomial data. *Journal of Econometrics 162*(2), 170–188.

Lin, X. (2018). Snap and food consumption among the elderly: a collective household approach with homescan data. Working paper, National Bureau of Economic Research.

Manski, C. F. (1987). Semiparametric analysis of random effects linear models from binary panel data. *Econometrica: Journal of the Econometric Society*, 357–362.

Matzkin, R. L. (2003). Nonparametric estimation of nonadditive random functions. *Econometrica 71*(5), 1339–1375.

Murtazashvili, I. and J. M. Wooldridge (2008). Fixed effects instrumental variables estimation in correlated random coefficient panel data models. *Journal of Econometrics 142*(1), 539–552.

Nevo, A. (2011). Empirical models of consumer behavior. *Annual Review of Economics 3*(1), 51–75.

Nickel, S. (1981). Biases in dynamic models with fixed effects. *Econometrica 49*, 1417–1426.

O'Donoghue, T. and M. Rabin (2003). Studying optimal paternalism, illustrated by a model of sin taxes. *American Economic Review 93*(2), 186–191.

Oliveira-Castro, J. M., G. R. Foxall, and T. C. Schrezenmaier (2006). Consumer brand choice: Individual and group analyses of demand elasticity. *Journal of the Experimental Analysis of Behavior 85*(2), 147–166.

Pacheco, J. and O. Valencia (2003). Design of hyprids for the minimum sum-of-squares clustering problem. *Computational Statistics and Data Analysis 43*(2), 234–248.

Peña, V. H., T. L. Lai, and Q.-M. Shao (2008). *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media.

Pesendorfer, M. (2002). Retail sales: A study of pricing behavior in supermarkets. *The Journal of Business 75*(1), 33–66.

Petrin, A. and K. Train (2010). A control function approach to endogeneity in consumer choice models. *Journal of marketing research 47*(1), 3–13.

Prifti, E. and D. Vuri (2013). Employment protection and fertility: Evidence from the 1990 Italian reform. *Labour Economics 23*, 77–88.

Puhani, P. A. (2012). The treatment effect, the cross difference, and the interaction term in nonlinear difference-in-differences models. *Economics Letters 115*(1), 85–87.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, Volume 4, pp. 321–333.

Rider, J., P. Berck, and S. B. Villas-Boas (2012). Eating healthy in lean times: The relationship between unemployment and grocery purchasing patterns.

Robbins, J. M., V. Vaccarino, H. Zhang, and S. V. Kasl (2001). Socioeconomic status and type 2 diabetes in african american and non-hispanic white women and men: evidence from the third national health and nutrition examination survey. *American journal of public health 91*(1), 76.

Schönberg, U. and J. Ludsteck (2014). Expansions in maternity leave coverage and mothers labor market outcomes after childbirth. *Journal of Labor Economics 32*(3), 469–505.

Seiler, S., A. Tuchman, and S. Yao (2019). The impact of soda taxes: Pass-through, tax avoidance, and nutritional effects.

Shocker, A. D., M. Ben-Akiva, B. Boccara, and P. Nedungadi (1991). Consideration set influences on consumer decision-making and choice: Issues, models, and suggestions. *Marketing letters 2*(3), 181–197.

Silver, L. D., S. W. Ng, S. Ryan-Ibarra, L. S. Taillie, M. Induni, D. R. Miles, J. M. Poti, and B. M. Popkin (2017). Changes in prices, sales, consumer spending, and beverage consumption one year after a tax on sugar-sweetened beverages in berkeley, california, us: A before-and-after study. *PLoS medicine 14*(4), e1002283.

Song, I. and P. K. Chintagunta (2007). A discrete–continuous model for multicategory purchase behavior of households. *Journal of Marketing Research 44*(4), 595–612.

Staubli, S. (2011). The impact of stricter criteria for disability insurance on labor force participation. *Journal of Public Economics 95*(9), 1223–1235.

Stock, J. and M. Watson (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association 97*, 1167–1179.

Stoker, T. (1993). Empirical approaches to the problem of aggregation over individuals. *Journal of Economic Literature 31*(4), 1827–1874.

Stroebel, J. and J. Vavra (2019). House prices, local demand, and retail prices. *Journal of Political Economy 127*(3), 1391–1436.

Sturm, R., L. Powell, J. Chirqui, and F. Chaloupka (2010). Soda taxes, soft drink consumption, and childrens body mass index. *Health Affairs 29*(5), 1052–1058.

Su, L. and G. Ju (2017). Identifying latent grouped patterns in panel data models with interactive fixed effects. *Journal of Econometrics*.

Su, L., Z. Shi, and P. Phillips (2016). Identifying latent structures in panel data. *Econometrica 84*(6), 2215–2264.

Su, L., X. Wang, and S. Jin (2017). Sieve estimation of time-varying panel data models with latent structures. *Journal of Business & Economic Statistics*, 1–16.

Su, L., X. Wang, and S. Jin (2019). Sieve estimation of time-varying panel data models with latent structures. *Journal of Business & Economic Statistics 37*(2), 334–349.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Ser. B 58*, 267–288.

Train, K. E., D. L. McFadden, and M. Ben-Akiva (1987). The demand for local telephone service: A fully discrete model of residential calling patterns and service choices. *The RAND Journal of Economics*, 109–123.

United States Department of Agriculture (2015-2020). Dietary guidelines for americans. (8).

Wooldridge, J. M. (2005). Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models. *Review of Economics and Statistics 87*(2), 385–390.

World Health Organization (2015). *Guideline: sugars intake for adults and children.* World Health Organization.

Young, A. (2019). Consistency without inference: Instrumental variables in practical application.